Preface

> *A man's life is merely a collection of events, building one upon the other. When all the events are tallied; the triumphs; the failures; the mistakes, their sum makes up the man.*
>
> — Neal Cassady
>
> *The Last Time I Committed Suicide, 1997*

The work presented in this thesis was performed at Parsimonious Modelling (PM) research group at the Helsinki Institute for Information Technology (HIIT), Department of Information and Computer Science (ICS) in Aalto University School of Science (AaltoSCI). I am deeply indebted to my instructor D.Sc.(Tech.) Jaakko Hollmén for his enthusiastic engagement in my research and his invaluable streams of advice, ideas, and support ranging from the minute technical details to the overall research ideas, as well as living a life. I am also grateful to Prof. Samuel Kaski for his supervision of the thesis.

I am thankful to Helsinki Doctoral Programme in Computer Science — Advanced Computing and Intelligent Systems (Hecse) for funding the last three years of my research. I am also thankful to HIIT and Finnish Centre of Excellence for Algorithmic Data Analysis Research (ALGODAN) for duration 2008–2013 which is itself funded by the Academy of Finland. They funded the first year of my graduate studies and together with Hecse, they also funded numerous conference trips and research visits in Finland and abroad. The research visits and conference trips have given me invaluable experiences in understanding research challenges and knowledge of the state of the art in my research area. It also provided me an opportunity to communicate with researchers both formally and informally helping me to network and collaborate.

A big share of thanks also goes to my colleagues in the PM research group (Mikko, Janne) and the whole ICS, HIIT, and ALGODAN including the support staffs for providing splendid working and research environment. I would like to thank everyone involved in teaching and organising the courses I attended for my graduate, and postgraduate degree. The courses have helped me prepare for my research and understand the state of the art in my field of research.

Collaboration with fellow researchers in the field has given me a deeper understanding in teamwork, work ethics, and work environment. Moreover, I have gained ideas from discussion with Prof. Nada Lavrač, Anže Vavpetič, Jan Kralj, Bimal Babu Upadhyaya, and Chen Meng. I thank the pre–examiners Prof. Dr. Marko Bohanec and Prof. Olli Yli–Harja for taking their precious time to review this thesis and providing invaluable comments to improve the thesis. I am grateful to Asst. Prof. Jeroen de Ridder, for the honour of having him as the opponent for my dissertation.

Outside of work, I would also like to thank my mates Gautam, Sandeep, Roshan, Bishal, and Subash. If it weren't for you guys, I might have graduated sooner but without you, I would have never graduated. My sincere gratitude also goes to my parents, family, and relatives for their everlasting encouragement and motivation. Last but not the least, I thank my wife, Manju, for being by my side, not only for the fun times, but also the ups and downs, and for being my support and my inspiration in all aspects of my life.

Espoo, March 17, 2016,

Prem Raj Adhikari

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Prem Raj Adhikari, Jaakko Hollmén. Patterns from Multiresolution 0–1 data. In *Jilles Vreeken, Nikolaj Tatti, and Bart Goethals, Editors, UP '10, ACM SIGKDD Workshop on Useful Patterns*, Washington DC, ACM, New York, NY, USA, Pages 8–16, July 25, 2010, DOI: 10.1007/s10844-013-0282-3, July 2010.

II Prem Raj Adhikari, Jaakko Hollmén. Fast Progressive Training of Mixture Models for Model Selection. *Journal of Intelligent Information Systems*, IN PRESS, Springer, DOI: 10.1007/s10844-013-0282-3, Published Online: December 2013.

III Prem Raj Adhikari, Jaakko Hollmén. Multiresolution Mixture Modeling using Merging of Mixture Components. In *Proceedings of Fourth Asian Conference on Machine Learning (ACML 2012)*, In Steven C.H. Hoi and Wray Buntine Editors, Volume 25 of Journal of Machine Learning Research—Proceedings Track, pages 17–32, November 4–6, 2012, Singapore, URL: http://jmlr.csail.mit.edu/proceedings/papers/v25/adhikari12.html, November 2012.

IV Prem Raj Adhikari, Jaakko Hollmén. Mixture Models from Multiresolution 0–1 Data. In *Proceedings of Sixteenth International Conference on Discovery Science (DS 2013)*, Johannes Fürnkranz, Eyke Hüllermeier, and Tomoyuki Higuchi, Editors, Volume 8140 of Lecture Notes in Com-

puter Science, Springer–Verlag, Berlin Heidelberg, pages 1–16, October 6–9, 2013, Singapore. DOI: 10.1007/978-3-642-40897-7_1 , October 2013.

V Prem Raj Adhikari, Anže Vavpetič, Jan Kralj, Nada Lavrač, Jaakko Hollmén. Explaining mixture models through semantic pattern mining and banded matrix visualization. In *Proceedings of Seventeenth International Conference on Discovery Science (DS 2014)*, Sašo Džeroski, Panče Panov, Dragi Kocev, Ljupčo Todorovski, Editors, Volume 8777 of Lecture Notes in Computer Science, Springer International Publishing Switzerland 2014, pages 1-12, October 8–10, 2014, Bled, Slovenia. DOI: 10.1007/978-3-319-11812-3_1, October 2014.

Author's Contribution

Publication I: "Patterns from Multiresolution 0–1 data"

Generally, mixture models and pattern mining algorithms can handle only single resolution data in their standard form. We propose different deterministic data transformation methods to transform datasets across different resolutions facilitating the integration of datasets. The integrated datasets are in single resolution. We then use pattern mining algorithms such as the maximal frequent itemset and probabilistic modelling methods such as mixture models to identify and compare the patterns and performance of the algorithms in different resolutions of data.

Forming of the original idea and designing of the methodology for the research are performed jointly by the authors. The current author implemented and performed all the experiments and wrote most of the manuscript. The second author suggested the corrections to the manuscript. The current author also presented the contribution at the conference.

Publication II: "Fast Progressive Training of Mixture Models for Model Selection"

Expectation Maximisation (EM) algorithm is a popular algorithm to learn the maximum likelihood parameters of the mixture model. However, EM algorithm requires apriori knowledge of the number of component distributions in the mixture model to learn the maximum likelihood parameters of the mixture model. This is often unknown apriori in most situations. In this publication, we propose an algorithm to efficiently train a series of mixture models each with different number of mixture components suitable for comparisons during model selection.

The authors are jointly responsible for the original idea of the contribution. The current author performed all the experiments and wrote most of the manuscript. The second author suggested corrections on the manuscript. The current author also presented an earlier version of this contribution [2] in a conference.

Publication III: "Multiresolution Mixture Modeling using Merging of Mixture Components"

In this contribution, we propose an algorithm to model multiresolution data by merging the similar components from different mixture models in different resolutions. The mixture models are generated in each data resolution separately but they incorporate the information from the data in other resolutions.

The current author is responsible for forming the original idea, and methodology of the work. The current author also performed the all experiments and wrote most of the manuscript. The second author provided useful suggestions and corrections to the manuscript. The current author also presented the research in the conference.

Publication IV: "Mixture Models from Multiresolution 0–1 Data"

In this contribution, we propose a multiresolution mixture model consisting of multiresolution mixture components. The structure of multiresolution mixture components are determined from the domain ontology which is known apriori. The individual mixture components provide the functionality of Bayesian networks.

The authors are jointly responsible for the original idea and designing the methodology for the research. The current author performed the experiments and wrote most of the manuscript. The second author suggested corrections to the manuscript. The current author also presented the contribution in the conference.

Publication V: "Explaining mixture models through semantic pattern mining and banded matrix visualization"

In this contribution, we propose three part exploratory approach to analyse multiresolution data. The three parts consist of clustering using mixture model, extracting rules from clusters using semantic data mining, and simultaneous visualisation of the clusters from mixture models and the rules from semantic data mining algorithm using banded matrices. The semi–automated methodology proposed in the contribution aims to provide exhaustive analysis of a complex real world multiresolution data.

The authors are jointly responsible for the idea and designing of the research methodology. The current author implemented the clustering part of the three part explanatory process. Writing the paper was a collaborative effort of all the authors.

List of Abbreviations

3Vs	Volume, Velocity, and Variety
aCGH	array Comparative Genomic Hybridization
AIC	Akaike Information Criterion
BAC	Bacterial Artificial Chromosome
BIC	Bayesian Information Criterion
cDNA	complementary Deoxyribonucleic Acid
CGH	Comparative Genomic Hybridization
CNV	Copy Number Variation
CPD	Conditional Probability Distribution
DNA	Deoxyribonucleic Acid
EB	exabytes
EM	Expectation Maximization
E–Step	Expectation Step
FISH	Fluorescence In Situ Hybridization
FP	False Positives
G–banding	Giemsa banding
GMM	Gaussian Mixture Model
GrC	Granular Computing
ICL	Integrated Classification Likelihood
IID	Independent and Identically Distributed
ISCN	International System for Human Cytogenetic Nomenclature
kbp	Kilo base Pairs
KL	Kullback Leibler
LHC	Large Hadron Collider
MAFIA	MAximal Frequent Itemset Algorithm
MAP	Maximum a Posteriori Probability
Mbp	Mega base Pairs

MCMC	Markov Chain Monte Carlo
MDL	Minimum Description Length
MGMM	Multiresolution Gaussian Mixture Model
MLE	Maximum Likelihood Estimate
MM	Malignant pleural Mesothelioma
MPSS	Massive Parallel Signature Sequencing
mRNA	Messenger Ribonucleic Acid
M–Step	Maximization Step
NGS	Next Generation Sequencing
PCR	Polymerase Chain Reaction
SOM	Self–Organizing Maps
TMA	Tissue Microarrays
TS–SOM	Tree Structured Self–Organizing Maps
TP	True Positives
WHO	World Health Organization

CHAPTER 1

INTRODUCTION

> *Data does not equal information; information does not equal knowledge; and, most importantly of all, knowledge does not equal wisdom. We have oceans of data, rivers of information, small puddles of knowledge, and the odd drop of wisdom.*
>
> — HENRY NIX
>
> *Keynote address, AURISA, 1990*

Synopsis

This chapter conceptualises the topic of this dissertation with respect to the methodology and application. The chapter also covers the motivations for research, contributions of the thesis to the scientific community, and organisation of the chapters of the thesis.

1.1 Data Explosion

Dictionary definition of data is a piece of information that ranges from the values or measurements of quantitative and qualitative variables to the description of objects or phenomenon [37]. In computing terms, data is any digitally stored information. Throughout history, data was universal, and found everywhere. However, only employees generated data in computing terms by keying in the handwritten information. Nowadays, users generate data on their own, for example, social network statuses or photos, thereby increasing the amount of data produced. Furthermore,

new machines such as automatic climatic conditions recorder and technologies such as Large Hadron Collider (LHC) produce colossal amount of data [104]. This astronomical increase in the amount of data is referred to as big data [104, 108]. Modern science revolves around the methods and ways to analyse the data generated in their field to stimulate scientific discoveries.

Production of data these days is such humongous that it surpasses the estimates of Moore's Law [123]. For example, 5 exabytes (EB) (1 EB= 10^{18} bytes = 1 billion gigabytes) of data was generated from the dawn of civilisation until 2003. Today, we create 5EB of data every two days [140]. Three properties: Volume, Velocity, and Variety (often referred to by 3Vs) define the big data. The volume of data and speed at which they arrive and leave the real time systems provide challenges in big data analysis. In addition, variety in the collected data also poses considerable challenges to research in big data.

Over the years, measurement technology has progressed enormously, and produces variety of data in addition to the large volumes of data because each cycle of improvement in measurement technology produces data in a different representation. The variety is the aspect of big data that is closest to the topic of this thesis. Nowadays, individual dataset in the sets of datasets often have higher dimensionality, d, than the number of samples, N, i.e., $d \gg N$. Therefore, challenge in big data analysis is large temporal, and/or spatial data dimensions which results in the curse of dimensionality [17]. Traditional algorithms succumb to the challenges posed by small sample high dimensional datasets. Therefore, it is imperative to develop novel methods to analyse multiple datasets, i.e., sets of datasets in different representations within a single analysis.

Biology is one of the largest producer of big data which necessitates novel computational methods to analyse such wealth of data and to convert data to knowledge and wisdom [74, 111]. There are varieties of biological phenomena often interlinked with one another making variety aspect of big data prevalent in biological data source. This tremendous increase in biological data coupled with the variety is impossible to interpret using visual analysis. Instead, it requires novel computational methods for thorough understanding of the biological phenomenon. The growth of biological data has produced both opportunities and challenges for researchers to develop algorithms and analysis methods in computational domain to extract biological meaning from vast amounts of data.

1.2 Machine Learning and Data Mining

Machine Learning is a core sub–area of artificial intelligence that intersects the discipline of computer science, and statistics. The aim of machine learning is to develop algorithms that learn from the observed data, and use the experience to improve the performance [9, 23, 68, 120]. Machine learning includes a myriad of statistical, probabilistic and optimisation, and induction algorithms that are applicable in different tasks such as classification, regression, clustering, and pattern discovery. Data mining, also known as knowledge discovery, is the process of extracting useful information such as patterns, from unstructured and enormous sets of data by analysing data from different perspectives [67].

Machine learning and data mining complement each other and it is difficult to make a clear distinction between the two. Nonetheless, machine learning algorithms are often used in the data mining process. Machine learning and data mining, although a new discipline, has a large active research community. The community has already developed a cohort of fascinating algorithms and methods to treat the concept classes, and elegant and clever ways to search through databases. Hence, machine learning and data mining methods can address the challenges posed by data intensive disciplines such as biology.

In application areas such as biology, the number of training samples are often limited even in the age of big data. In contrast, the data dimensionality increases considerably. For example, in genetics, number of cancer patients is constant while the new technology can measure the finer units of the phenomenon generating data with large dimensionality. The implication of increasing dimensionality is that, with a limited size of training samples, the performance of the algorithm deteriorates as the number of features increases. This phenomenon is also called Hughes phenomena, or Hughes effect [76] or more generally as a curse of dimensionality [17].

1.2.1 Mixture Model

A mixture model is a probabilistic modelling technique in machine learning and data mining community which models a data distribution under the assumption that all the data points are generated from a mixture of parametric probability distributions [23, 45, 115]. Apart from this assumption of data origination, mixture models are flexible probabilistic models with varying uses such as model based clustering, classification,

image analysis, and collaborative filtering in analysis of high dimensional data. Mixture models are suitable for the choice of any probability distributions such as the Gaussian, Bernoulli, Poisson, and Dirichlet. In this thesis, mixture models analyse multiresolution data probabilistic clustering setting. Chapter 3 discusses mixture models in detail.

1.3 Challenges of Multiresolution Data

Measurement of physical phenomenon such as distance, weight, and time started since the time immemorial and has been the cornerstone of our knowledge and learning [92]. Measurement has also become integral part of our everyday life. The inventions and discoveries of the modern world would cease to exist in absence of measurement technology. The measurement technology has been continuously improving over the years. Result of a measurement process is generation of the data. The older generation technologies measure only the coarser unit of the phenomenon generating data in coarse resolution. In contrast, the newer generation technologies measure the finer units of the phenomenon producing the data in fine resolution [44, 54, 110]. Resolution here defines the amount of information in each data sample, i.e., the level of detail.

Multiresolution data is generated when the same phenomenon is measured in different levels of detail [11, 54, 165]. Thus, the multiresolution data describes the same phenomenon in different data representations. Different data representation is a broader challenge in machine learning and data mining community where datasets are represented in different forms such as audio, video, image, table, and text. This thesis concentrates on different data representation only in the context of dimensionality, i.e., datasets are same except for the data dimensionality. Nevertheless, the proposed algorithms and methods can possibly be extended to other different data representations. The measurement of time is one of the simplest illustrations of multiresolution data. We can measure time in fine units such as seconds and minutes producing data at a fine resolution. In contrast, we can also measure time in coarse units such as months, and years producing data in coarse resolution.

Multiresolution data often forms a part of hierarchy as shown in Figure 1.1. For example, the world is a collection of different continents such as Asia, Europe, and Africa. This generates coarser view of data. Similarly, the world is also a collection of different countries such as Singa-

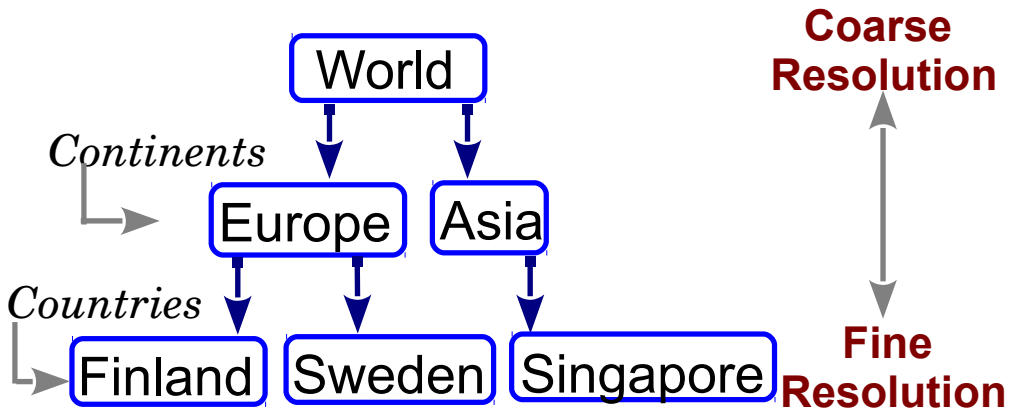

Figure 1.1. Example of part of hierarchy in real world scenario. The figure shows the geographical division as the part of hierarchy which when measured results in multiresolution data. The world is divided into continents and continents into countries.

pore, Finland, and Sweden. These countries can be grouped into different continents. Furthermore, these countries can be divided into municipalities, and the municipalities into streets, and the streets into blocks. This hierarchy forms a multiresolution data which represents a part of hierarchy [139]. This division of the world is just an illustrative example, as the sources of multiresolution data are varied, for example, telecommunications, hydrology, and biology [165]. Chapter 2 discusses multiresolution biological data used in the experiments of the thesis.

1.4 Contributions of the Thesis

This thesis addresses an important challenge encountered in data analysis: what should be done when the data to be analysed are represented differently. The thesis presents different frameworks and methods amalgamating probabilistic modelling and pattern mining domain. The presented methods handle irregular, and heterogeneous division of data in different representations. The major scientific contributions in this thesis are summarised in the following list.

- Different deterministic data transformation methods are proposed to transform the multiresolution datasets from one resolution to another. The transformed datasets in same resolution can be integrated and modelled in same resolution.

- A computationally efficient algorithm is proposed to train a series of mixture models to aid model selection. The trained mixture models in the series differ in number of components but are otherwise similar to each other. This provides an effective means to compare different model selection criteria such as likelihood, AIC, and BIC using different model selection techniques such as cross–validation.

- A mixture modelling solution is proposed to model multiresolution data by merging the mixture components of different mixture models in different resolutions. The proposed mixture modelling solution initially trains a mixture model in each resolution and merges the similar mixture components across different resolutions to incorporate information in multiple resolutions.

- An algorithm that uses domain ontology, known apriori, to determine multiresolution mixture components of the mixture model is proposed to build a single mixture model for multiresolution data. Each individual mixture component is a fully functional Bayesian network.

- A three part methodology is proposed to analyse the multiresolution data blending clustering using mixture models, pattern mining using semantic data mining, and visualisation using banded matrices.

1.5 Organisation of the Thesis

The thesis consists of two parts: an introductory part consisting of six different chapters and publications. In the introductory part, this chapter introduces the research domain, and the Chapter 2 introduces multiresolution data with a focus on cancer genomics, and reviews the previous work in multiresolution analysis and the related areas. Chapter 3 describes mixture models and model selection in mixture models. It also summarises our contribution for efficient training of a series of mixture models (Publication II).

Chapter 4 forms the crux of this thesis and discusses our contributions in multiresolution modelling. First, multiresolution data is modelled using deterministic data transformation methods for data integration (Publication I). Second, multiresolution data is modelled by merging the simi-

lar mixture components of different mixture models in different resolutions. The merging of mixture components models the interaction between the models in different data resolutions (Publication III). Third, a multiresolution mixture model having multiresolution mixture components is proposed to analyse the multiresolution data with a single mixture model. Structure of multiresolution components is known from the domain ontology (Publication IV). Finally, a comprehensive solution for the analysis of multiresolution data is provided using three part methodology comprising of clustering, semantic pattern mining, and banded matrices (Publication V). Chapter 6 summarises the findings, presents the conclusions of the research, and also outlines the possible future work related to the topic of the thesis.

CHAPTER 2

MULTIRESOLUTION DATA AND ANALYSIS METHODS

> *Data matures like wine, and the applications like fish.*
>
> — JAMES GOVERNOR
>
> *James Governor's Monkchips, 2007*

Synopsis

This chapter describes the application area and the dataset used in the experiments. The chapter also describes the usefulness of domain ontology in data analysis; and the multiresolution data in the domain of biology. Finally, the chapter also briefly reviews the literature and discusses the related areas of multiresolution modelling.

Human beings are diploid organisms having two homologous copies of each chromosome one each inherited from each parent. Copy Number Variations (CNVs) are structural variations in genome such that a region on the genome will have different copies of DNA [146]. In human beings, normal copy number is two because each child inherits one copy from each parent. Deletion or loss is the condition when the copy number is less than two. Duplication or gain is the condition when the copy number is more than two. Similarly, amplification is the condition when the copy number increases to more than 5. Some of the cancer patients have shown more than hundred copies [158]. There are other different kinds of variations but this thesis concentrates on copy number aberrations.

2.1 Chromosomal Aberrations in Cancer

Cancer is a heterogeneous collection of diseases characterised by abnormal and uncontrolled growth of cells; their ability to migrate to other parts of human body and destroy neighbouring cells and tissues [24]. Cancer rates have been increasing rapidly around the globe. Recent World Health Organisation (WHO) report showed that number of cancer patients escalated to 14.1 million in 2012, and cancer was responsible for 8.2 million deaths in 2012 [147]. The menace of cancer is increasing and WHO estimates that cancer will rise by 57% worldwide in the next 20 years signalling an imminent human disaster. The cost of cancer is also increasing rapidly. In 2010, estimated global cost of cancer reached approximately 963 billion euros [147], which is nine times more than the total budgeted expenditure of Finland.

A wide range of genetic mutations and molecular mechanisms known as chromosomal aberrations are identified as the hallmarks of disorders such as Cancer, Schizophrenia, and infertility [158]. In cancer research, identification and characterisation of chromosomal aberrations are crucial for studying and understanding pathogenesis of cancer. Moreover, study of chromosomal aberrations provides necessary information to select the optimal target for cancer therapy on individual level [91]. Study of chromosomal aberrations also has other clinical applications such as studying multiple congenital abnormalities and identifying the family history of Down syndrome [130].

2.2 Measurement Technology in Biology

Years of evolution and adaptation have made organisms complex biological beings [116]. Ever improving measurement technologies have also provided the facilities to measure the complex phenomena in biology [41]. After the discovery of DNA in 1953 [162], different measurement methods have been proposed to measure the genome. First sequence of lac operator of 24 bp was published twenty years after the discovery of DNA in 1973 [58]. Figure 2.1 summarises the evolution of DNA sequencing technology from the 1973. Initially, different banding methods such as G–banding and Q–banding technologies were developed to produce a visible karyotype by staining the chromosomes [19]. A karyotype here denotes the set of all chromosomes in an organism. Alongside the banding tech-

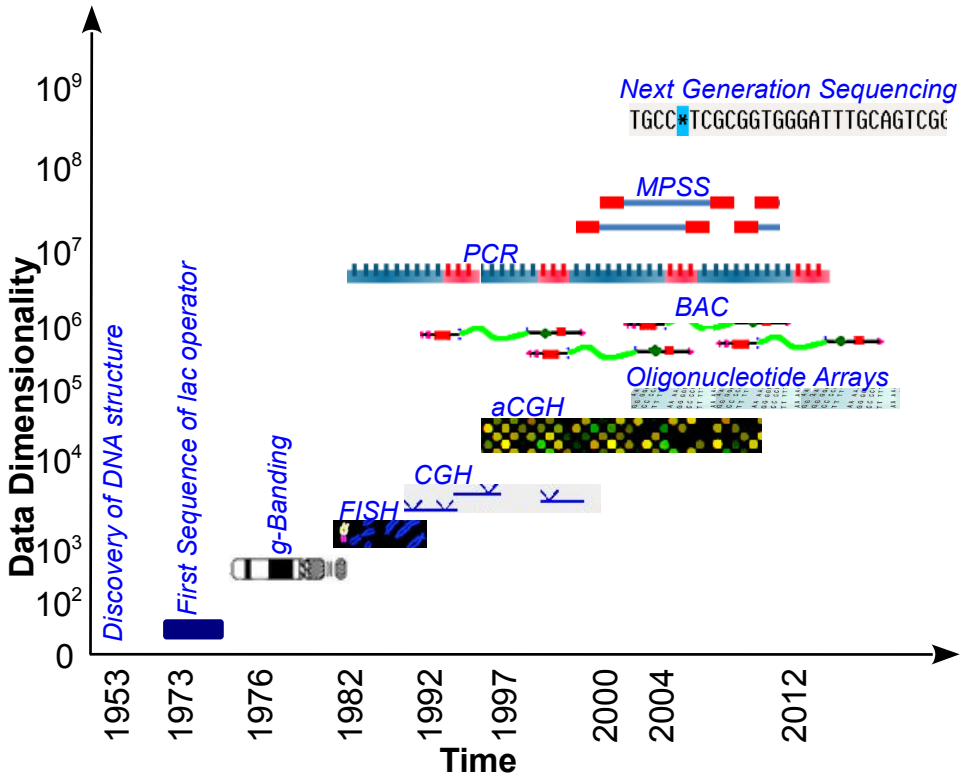

Figure 2.1. Evolution of measurement technology in biology described in terms of their level of detail in measurements and time of usage.

nology, FISH (Fluorescence In Situ Hybridisation) was developed to detect the presence or absence of DNA sequences on chromosomes. Similarly, microarray technologies such as the Comparative Genomic Hybridisation (CGH) [85] and array Comparative Genomic Hybridisation (aCGH) [134] were developed to study the Copy Number Variations (CNV) without requiring culturing of cells. Additionally, Bacterial Artificial Chromosome (BAC) was developed to sequence the genomes of organisms.

Similarly, Oligonucleotide arrays that uses oligos of short lengths (less than 25 bases) were developed to test large number of oligos in presence of smaller number of targets [103]. In addition, promoter arrays were developed to probe thousands of promoter sequences in one array experiment [161]. Besides, Massive Parallel Signature Sequencing (MPSS) was developed to analyse the level of gene expression by identifying and quantifying Messenger Ribonucleic Acid (mRNA) transcripts in the sample [25]. Likewise, Polymerase Chain Reaction (PCR) were developed to amplify one or small number of copies of DNA thereby generating large number of copies of particular DNA sequences useful for biomedical application such as DNA sequencing and diagnostic purposes [12].

Around the beginning of this century new technology known as next generation sequencing (NGS) had resounding impact in DNA sequencing. In [109] and [110], authors summarise the improvement in DNA sequencing which has positive impact on the biomedical research providing high throughput and high resolution techniques to explore, and answer genomewide biological questions. The Carlson curve accurately predicted the doubling time of DNA sequencing technologies measured in terms of cost and performance [27]. Furthermore, the curve illustrates the dramatic decrease in cost which is sometimes hyperexponential and similar dramatic improvements in technology to measure biological phenomenon such as DNA sequencing and synthesis, gene expressions, and protein structures.

These improvement in measurement technology in biology over the period of time produces data in different representation. Consequently, multiresolution data are also present in biology. For example, measurements from an older generation technology (eg. G–banding) can be represented in data with dimensionality in hundreds [19, 143]. In contrast, newer generation technology such as microarray measures the same karyotype generating the data of dimensionality of thousands [85, 134]. In addition, latest technology known as Next Generation Sequencing produces the data with millions of dimensions [109, 138].

The Figure 2.1 shows major changes in sequencing technology. However, within each generation of technology there are several minor improvements. For example, aCGH improves the mapping resolution of 20Mbp (Megabase Pairs) to 100 Kbp (Kilobase Pairs) over its predecessor CGH. Similar methods within a generation of technology also produce data in different resolutions because of improvements within the technology such as microarrays and banding. For example, authors in [155] use microarray data in two resolutions of 44000 and 244000 measurements per microarray measured by Agilent 44B and 244A aCGH platforms to classify different types of leukemias. Similarly, in NGS, different vendors have produced different sequencers for commercial use [102].

Studying the data generated by different technologies above produces wide range of benefits, especially in understanding of the biological phenomenon. Therefore, computational methods have been used to analyse the generated data. The phenomenon of doubling of number of transistors in a chip within 18 to 24 months, often known as Moore's Law [123], has improved the processing power of computers exponentially. Similarly,

with the advent of Internet and other communication technologies and protocols; communication systems have also improved dramatically. The data storage capacity is also rapidly rising. These advancements have resulted in improved computing power thus facilitating development of novel algorithms to analyse the generated data.

Pan–cancer Analysis

In addition to the data in different resolutions, efforts have been made to study different cancer types by collecting data from different sources in pan–cancer initiative [127]. The aim of the study is to develop an integrated picture of commonalities, differences, and emergent themes across tumour lineages. The initiative involves multiple datasets and multiple cancers showing possible utility of multiresolution methods in pan–cancer initiative. In the previous research of our research group, we have considered all cancers within a single analysis [125].

2.3 Multiresolution Chromosomal Amplification Data

Similar to the array technology and next generation sequencing, the International System for human Cytogenetic Nomenclature (ISCN) has defined five different resolutions of the chromosome namely: 300, 400, 550, 700, and 850 in G–banding [143]. Each resolution defines the precision in division of karyotype. For example, in coarse resolution, a karyotype is divided into 312 (≈ 300) different regions, i.e., with lower precision. In contrast, in fine resolution, a karyotype is divided into 862 (≈ 850) different regions, i.e., with higher precision compared to resolution 300.

Figure 2.2 shows five different resolutions in chromosome 21 according to the ISCN standard. The figure depicts the division of regions and chromosome nomenclature with an example in chromosome 21. Chromosome 21 is chosen for visualisation because it is the smallest chromosome. Chromosome 21 is divided into 8, 8, 10, 12, and 14 regions in resolution 300, 400, 550, 700, and 850. The nomenclature of the regions and their division in different resolutions are irregular and hierarchical [143]. Some regions are undivided whereas other regions are divided into different number of regions. For example, the regions 21p12 and 21p13 are undivided in all the resolutions where as the region 21q22 is divided into 3 and 5 different regions in resolution 550 and 850. This division of karyotype

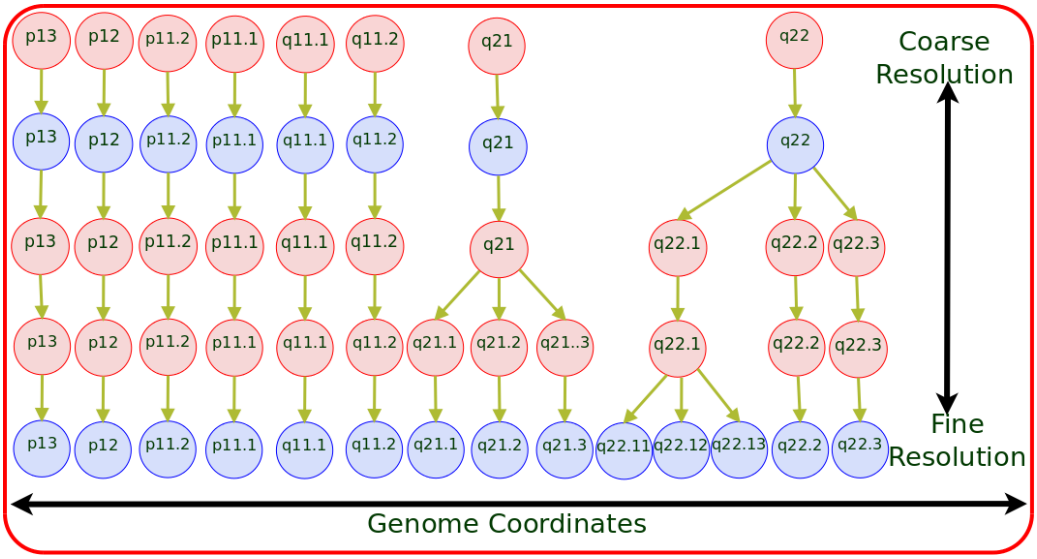

Figure 2.2. A typical relationship between multiple resolutions of genome. Figure shows chromosome 21 in five different resolutions of genome as defined by ISCN standard. The division is irregular, and hierarchical but consistent because of the ISCN standard. Chromosome 21 is chosen for the clarity of the presentation because it is the smallest chromosome. Y–axis denotes different resolutions of genome while x–axis denotes spatial coordinates (different regions) of the genome. Figure is adapted from Publication IV.

in different levels of detail allows measurement technologies to generate data in multiple resolutions. Each chromosomal region in coarse resolution is related with a chromosomal region in fine resolution with a one to many relationship. Given the measurements of same subject in two different resolutions, the aberrations should be consistent with each other, i.e., the aberrations should be the same except for some measurement errors.

For the experiments, two different datasets were available in coarse resolution and fine resolution. Researchers at the University of Helsinki compiled a dataset of chromosomal amplification in coarse resolution reading through the literature published between 1992–2002 [94]. All 838 journal articles were read through manually. The data describes the chromosomal amplification patterns of 4590 cancer patients in coarse resolution, i.e., resolution 400 where a karyotype is divided into 393 different regions. Similarly, data in fine resolution extracted from [13, 14] describes chromosomal amplification in fine resolution, i.e., resolution 850 where a karyotype is divided into 862 different regions. Since the cancer patients were not the same, there is no direct correspondence between data samples in two different resolutions. Therefore, most of the analysis methods discussed in this thesis consider unsupervised methods which learn the hidden structure in the data without the help of the class labels [23, 120]. If the measurements were available from the same cancer patients in two

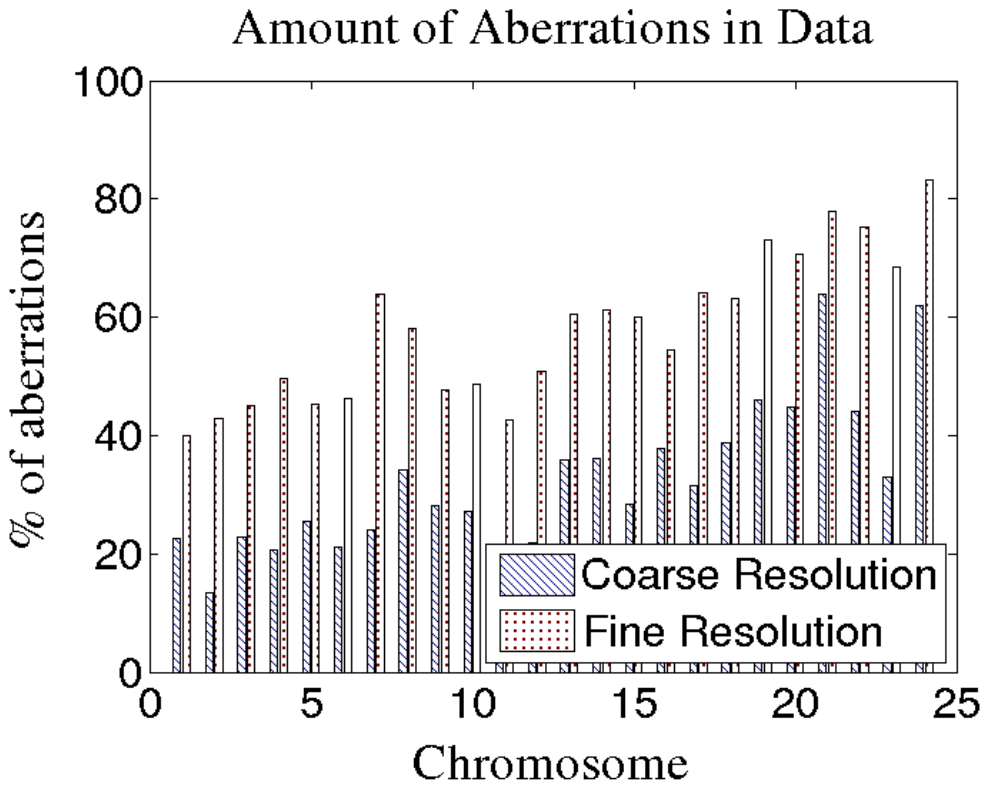

Figure 2.3. Amount of aberrations in each chromosome in two datasets in different data resolutions. The bar diagram shows that chromosomes in fine resolution are comparatively more aberrated than the coarse resolution.

different resolutions, we can expect consistent matching in aberrations except for measurement errors.

In the coarse resolution data, a total of 26527 (out of 1,803,870) matrix elements are aberrated which accounts for approximately 1.5% of the total matrix elements in the dataset. In all our experiments, we process the data chromosomewise to reduce the data dimensionality and with the expectation of finding chromosome specific patterns to describe different cancers. When the data is divided into each chromosome, there are some samples which do not show aberration in any of the chromosomal regions. Such data samples are deleted as we are interested in modelling chromosomal aberration patterns and their relation to cancer, not their absence. Therefore, number of samples and data dimensionality in each chromosome is different. We therefore calculate percentage of aberrations in each chromosome in each resolution for comparison purposes. Figure 2.3 depicts the amount of aberrations in both coarse resolution and fine resolution data. Data in fine resolution shows more aberration than the data in the coarse resolution. The percentage of aberrations are approximately 50% overall, while the minimum and maximum are approximately 15% and 80% respectively.

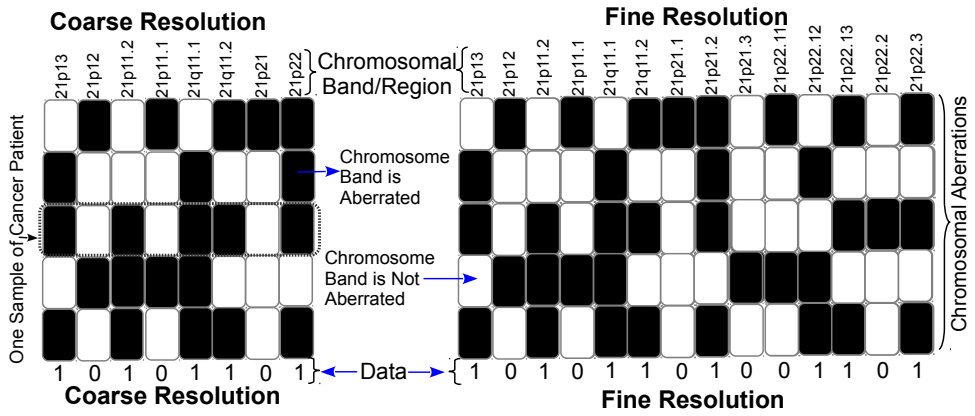

Figure 2.4. Visualisations of data describing chromosome 21 in two different resolutions: 300, and 850. Each sample, i.e., each row denotes one cancer patient and each column determines a chromosomal region. The black colour denotes presence of amplification and white col or denotes the absence of amplification. The two different panels in the figure depict the same phenomenon measured at different resolutions. Some chromosomal regions (variables or features in machine learning terms) such as 21p21 in left panel have been divided into different regions such as 21p21.1, 21p21.2, and 21p21.3 in the right panel. Figure is adapted from Publication II.

Figure 2.4 depicts five samples of data from chromosome 21 in both the coarse and the fine resolution. In the Figure 2.4, rows denote the cancer patients and the spatial coordinates on the X–axis denote the chromosomal region. In addition, white colour denotes value of zero (0), i.e., the absence of amplification, and black colour denotes the value of one (1), i.e., amplification in that specific region of genome for that specific cancer patient. The left panel of the Figure 2.4 shows that one region 21p21 in coarse resolution is divided into 3 regions in the fine resolution: 21p21.1, 21p21.2, and 21p21.3 as shown in the right panel of the figure. In contrast, some of the regions such as 21p13 and 21p12 are same in both coarse and fine resolution. Some regions are undivided while other regions are divided into varying number of regions. Nevertheless, the division is consistent because of the ISCN standard. Detailed description of the amplification dataset in coarse resolution can be found in [125].

2.4 Ontology of Multiresolution Data

The concept of ontology transcends back to the dates of noble philosophers Aristotle, Parmenides, and Jacob Lorhard, who used the term ontology in the philosophical context to describe the state of being, and reality [34]. Recently, the term ontology has found its prominence in computer and information science community. In computer science community, ontology

is the mechanism for explicit description of the conceptualisation of the knowledge represented in the knowledge base [63, 151].

Ontology is a popular methodology to describe the semantics of the data in machine learning and data mining community [132]. Recent studies have shown that relevant additional knowledge enhances the knowledge discovery process of empirical data [132]. Expansion of semantic web and increasing availability of domain knowledge as ontologies has resulted in growth of semantic data. Semantic data mining algorithms address the challenge of mining abundance of knowledge encoded in domain ontologies constrained by the heuristics computed from the empirical data [157].

Multiresolution data conceptualises one of the essential ontological dichotomies of universals and particulars in metaphysics [57, 139]. The data in the coarse resolution can be conceptualised as universals whereas data in fine resolution can be conceptualised as particulars. Therefore, we can use ontological information in modelling multiresolution data as in Publication IV and Publication V.

Biological systems are complex consisting of many interwoven subsystems that effect the functionalities of each other [89]. As a result, chromosomal amplifications can effect, and be effected by other biological phenomenon. Furthermore, cancer is a multifactorial disease and the heterogeneity of cancer also suggests that biological phenomenon besides chromosomal aberration can catalyse the development of cancer. For this reason, additional background knowledge in biology was used to enhance the comprehensive analysis of chromosomal amplification datasets and to help understand the phenomenon of cancer. The additional knowledge used in the analysis of multiresolution data are the taxonomy of hierarchy of chromosomal regions, the cancer genes, virus integration sites, fragile sites, and amplification hotspots in Publication V. Only taxonomy of hierarchy of regions is used as background ontology in Publication IV.

The mutations in genes resulting to a larger extent by “acquired mutation” and to a lesser extent by “germline mutation”, known as cancer genes, are one of the most prominent causes of cancer [49]. Authors in [49] have listed the cancer genes and compared them to the complete gene set revealed by the human genome sequence. Similarly, fragile sites are non-randomly distributed loci on human chromosome that show a constriction or a gap and increased frequency of chromosome breakage under the conditions of partial replication stress [42, 141]. The fragile sites are often found rearranged in cancers [60]. Virus integration sites are the loca-

tions in chromosome where the viral Deoxyribonucleic acid (DNA) inserts into host cell DNA [88]. Viruses are responsible for approximately 12% of cancers [88, 172]. Amplification hotspots are frequently amplified chromosomal locations in cancer patients identified using computational modelling in [125]. The semantic data mining methods use these additional knowledge to enhance the knowledge discovery process in Publication V in semantic subgroup discovery framework.

2.5 Pattern Mining

Pattern mining is a popular branch of data mining that aims to extract interesting, relevant, and meaningful patterns from the data [66, 67]. Frequent itemset mining is one of the first and most popular pattern mining algorithm. Itemsets are a set of items or columns in a 0–1 dataset having high concentration of 1s and are used as patterns in a 0–1 dataset [152]. Let $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_n$ be the attributes (items) of a dataset, $\mathcal{D}$, and σ be the given support. A frequent itemset is a set $\mathcal{F}$ of items of $\mathcal{D}$ such that at least a fraction of σ of the rows of $\mathcal{D}$ have 1 in all columns of $\mathcal{F}$ [4, 106].

Anti–monotone property of frequent itemset suggests that if an itemset is frequent, then all its subsets are also frequent [51]. Hence frequent itemsets generate a larger number of patterns making it difficult to report and interpret the results. Maximal frequent itemset ameliorates challenges posed by larger number of patterns in frequent itemsets. An itemset is maximal frequent if none of its immediate supersets is frequent [26]. We use maximal frequent itemset in Publication I to compare and report patterns across different resolutions.

Similarly, association rule is a popular data mining methodology to determine the interesting relations between variables based on different measures of interestingness [5, 72, 93, 133]. Most initial studies in association rule mining focused on finding interesting patterns from the large databases in an unsupervised setting. Nevertheless, association rules have been used in classification [82, 101]. Continuing with the research on association rules and classification, domain of subgroup discovery has emerged as a popular data mining methodology for labelled data. Subgroup discovery aims at finding interesting rules from the data that best describe the target variable [53, 71, 129]. Additionally, contrast set mining aims to learn the variables that differentiates one group of target variables from the rest, i.e., the most discriminating sets of variables [16, 129].

Semantic data mining method is a branch of data mining that uses taxonomies and ontologies of background data to improve the performance of algorithms [98, 156, 157]. Semantic data mining has recently gained research interest in pattern mining community because of the availability of large amount of data in the form ontologies encoded in semantic web [98]. Especially, the additional knowledge are abundantly available in biology as discussed in Section 2.4. In Publication V, we use semantic data mining algorithm to explain the clustering results obtained by probabilistic clustering using background knowledge discussed in Section 2.4.

2.6 Related Work in Multiresolution Data Analysis

Multiresolution analysis and modelling research community is growing steadily because of the pragmatic approach in dealing with datasets in different representation within a single analysis and also because of the increasing availability of multiresolution data in different application areas [11, 69, 80]. For instance, authors in [136] have improved the efficiency of boosting algorithms in regression and classification, using the model–driven and data–driven multiresolution strategy. Similarly, multiresolution trees have been used for object recognition in homogeneous data based on recursive neural networks [21]. In addition, multiresolution visualisations have been used to visualise large volumes of complex data using semantic analysis to infer increasing levels of meaning from the data [79]. Similarly, tree structured self–organising maps (TS–SOM) have been proposed in the literature as a multiresolution representation of several self–organising maps (SOMs) [95].

Multiresolution Probabilistic Models

Multiresolution modelling has also received research interest in probabilistic modelling domain. Most of the focus in this thesis has been the use of probabilistic models, namely mixture models, to analyse multiresolution data. Traditional machine learning and data mining methods, such as mixture models, are unable to analyse multiresolution data in their standard form because of the difference in representations of data in different resolutions. The only approach to model multiresolution data is to model each resolution separately and at best compare the results. Nevertheless, multiresolution models have found their usage in the literature,

especially, in the image processing domain. For example, multiresolution kd–trees have been used to improve the performance of mixture models and reduce the cost associated with the Expectation Maximisation (EM) algorithm [122]. Similarly, multiresolution kd–trees have also been used to build robust models against the outliers using the EM algorithm [128]. Similarly, multiresolution binary trees have been used to store probability values efficiently both in terms of time and memory [18].

Authors in [124] have improved the performance of Gaussian Mixture Model (GMM) using wavelet subbands with an additional feature of incorporating variable number of components in the GMM. The GMM in [124] can use any multiresolution based decomposition for background suppression. Authors in [166] show that Multiresolution Gaussian Mixture Model (MGMM) adapts to smooth motions. The authors then apply the MGMM to estimate the visual motion. Similarly, authors in [117] propose efficient algorithms to learn a mixture of trees model in a maximum likelihood and Bayesian network framework for discrete multidimensional domains.

Related Areas

Multiresolution analysis and modelling shares commonality with various research areas and applications. The following sections briefly review the work on multiresolution modelling in the relevant research areas.

Multiscale Analysis and Scale space Theory

Multiresolution modelling is often synonymously used in literature with the scale space theory [99] and also multiscale analysis [163]. In image processing domain, pyramid structures generated after successive smoothing, and subsampling produces a multiscale representation [99]. Similarly, in scale space theory a scale parameter, t, handles images at different scales. Scale space representation, an improvement over multiscale representation, has an ability to compute representation at a desired scale. Authors in [8] address an important challenge in cancer research by identifying densely connected components of known and putatively novel cancer genes in protein protein interaction networks using a multiscale diffusion kernel. The results in [8] demonstrate the importance of multiscale analysis as the putative cancer genes and network are significant at different diffusion scales. Similarly, authors in [38] detect statistically significant co–mutations in multiple independent insertional mutagenesis screens. The significance is estimated on multiple scales and results

are visualised in scale space thus providing valuable supplementary information on the putative cooperation. Multiscale analysis and scale space theory also provide functionalities to address the challenges of image representation at different resolutions. Similarly, a family of methods known as super–resolution has been used to increase the resolution of images and videos [119]. Generally, both multiscale and scale space methods work in model domain. However, multiresolution methods developed in this thesis are the result of multiresolution challenges arising in the data domain.

Wavelets

Wavelets are appropriate methods to describe the mathematical phenomenon such as functions and signals at different levels of resolution [105]. Wavelet analysis have been popular tool in multiresolution analysis [81]. However, the classical wavelets based techniques are useful in regular, consistent, and homogeneous setting. Hence, wavelets cannot directly handle the irregularities in the chromosomal amplification data.

Learning from Multiple Sources

Similar to multiresolution modelling, learning from multiple sources aims to ameliorate the problem of curse of dimensionality, or Hughes effect by exploiting any related additional datasets such as earlier measurement experiments [36]. Unlike multiresolution modelling, the additional datasets may only be weakly related to the analysed dataset. The paradigm of learning from multiple sources is extended to the paradigms of multiview [150], multiway [78], and multitask learning [29].

Data Fusion

The domain of data fusion shares a common ground with the domain of multiresolution modelling. Data fusion integrates multiple data and knowledge depicting the same real world phenomenon in a single, logical, precise, and useful knowledge base [61]. Data fusion techniques are often used to combine data from multiple sensors in such a way that the inference from the combined data is better than that from individual sensors. Data integration approaches have also been widely used in bioinformatics domain. For example, authors in [62] have proposed integrated database and software system that enables retrieval and visualisation of biological relationships across heterogeneous data sources. Similarly, authors in [87] combined data from complementary Deoxyribonucleic Acid (cDNA) arrays and tissue microarrays (TMA) to study the molecular changes in

malignant pleural mesothelioma (MM). The study shows that novel proteins associated with cell adhesion are expressed either directly or as a regulatory mechanism in MM. The process of data fusion takes place at the different stages of analysis but it is a common practice to merge the data at the earliest stage of analysis in a single resolution. Data fusion techniques have also been used in multiresolution analysis, especially in remote sensing [28].

Granular Computing

Granular computing (GrC) has roots in multiresolution modelling [10]. GrC is a multidisciplinary field of study comprising of theories, methodologies, and tools to analyse data using the granules in data [170]. Granular computing aims to divide data into different intrinsic resolutions to solve a problem which resembles with multiresolution modelling framework.

CHAPTER 3

MIXTURE MODELS AND MODEL SELECTION

> *The purpose of models is not to fit the data but to sharpen the questions.*
>
> — SAMUEL KARLIN
>
> 11^{th} *R A Fisher Memorial Lecture (1983)*

Synopsis

This chapter introduces mathematical foundation and formulation of mixture models. The chapter also discusses the model selection in mixture models. This chapter also discusses one of the associated publications where we propose a computationally efficient algorithm to train a series of mixture models to aid model selection procedure.

Classical probability distributions such as Gaussian, Bernoulli, and Poison provide methods for probabilistic modelling of data [160]. However, in the real world scenario, a single probability distribution cannot emulate the complexity in the data. Nevertheless, a combination of sufficiently large number of probability distributions can possibly emulate complexity in the data. Such combination of multiple classical probability distributions forms a mixture model. Formally, mixture models are semiparametric latent variable models that model a complex data distribution by weighted sum of different probability distributions [23, 45, 115].

The probability distributions within a mixture model, known as component distributions, describe the observations present in the data. The formulation of mixture model involves determining the number of compo-

nents in the mixture model, their associated distribution, and identification of the component generating the specific data sample [115]. Mixture models are often used in hard clustering analysis as in this thesis. In hard clustering, only one component is responsible to generate a specific data sample. Mixture models also provide the option of learning soft clustering. In soft clustering, a data sample belongs to more than one cluster with a certain degree of association [23]. A standard formulation of the mixture model assumes that the samples are independent and identically distributed (IID). Under the assumption that data originates from a known number of components, J, the probability density of a mixture model can be expressed as the weighted sum of its component distributions as:

$$p(x) = \sum_{j=1}^{\mathrm{J}} \pi_j P_j(x \mid \boldsymbol{\theta}_j), \tag{3.1}$$

where j indexes the component distributions. In the Equation (3.1), the mixing proportion (mixing or mixture coefficient) is denoted by π_j for the j^{th} component in the mixture model. It determines the weight of the component distribution in the overall mixture model. Mixing proportions satisfy the property of convex combination such that $\int p(x)dx = 1$, $\pi_j > 0$, and $\sum_{j=1}^{J} \pi_j = 1$ [45]. Similarly, the parameters $\boldsymbol{\theta}_j$ in Equation (3.1) denotes the parameters of the j^{th} component distribution of the mixture model. Application area dictates the choice of distributions, which in literature is dominated by the distributions from exponential family such as Gaussian, and Dirichlet [115]. In this thesis, Bernoulli distribution is the preferred distribution because the datasets are 0–1 datasets describing chromosomal amplifications.

3.1 Mixture Models of 0–1 Data

Finite mixture model of multivariate Bernoulli distributions for a dataset, $\boldsymbol{X}$, of dimensionality, d, are parametrized by $\boldsymbol{\Theta} = \{J, \{\pi_j, \Theta_j\}_{j=1}^{J}\}$. The dataset, $\boldsymbol{X}$, consists of samples $\mathrm{x}_1, \ldots, \mathrm{x}_N$ in such a way that $\boldsymbol{X} = \{\mathrm{x}_1, \ldots, \mathrm{x}_N\}$. Replacing the general probability distribution function with the distribution of choice, i.e., Bernoulli distribution, a mixture model of multivariate Bernoulli distribution can be mathematically expressed as [45, 167]:

$$p(\mathbf{x} \mid \mathbf{\Theta}) = \sum_{j=1}^{\mathrm{J}} \pi_j \prod_{i=1}^{\mathrm{d}} \theta_{ji}^{x_i} (1 - \theta_{ji})^{1-x_i}, \tag{3.2}$$

where j indexes the components, and i indexes the data dimensionality. x_i denotes the data point such that $x_i \in \{0, 1\}$. The parameter of a random variable θ_{ji} denotes the probability of the variable taking the value 1 in i^{th} dimension of the j^{th} component. We can collect all the random variables in a component in a vector, Θ_j such that $\Theta_j = [\theta_{j1}, \theta_{j2}, \theta_{j3}, \ldots, \theta_{jd}]$. Similarly, we can collect all the parameters of the mixture model including mixture components in a matrix, $\mathbf{\Theta}$ such that $\mathbf{\Theta} = \{J, \{\pi_j, \Theta_j\}_{j=1}^{J}\}$. The parameter values that maximise the log–likelihood function of the parameters can be defined using maximum likelihood principal [23] as:

$$\mathcal{L}(\mathbf{\Theta} \mid \boldsymbol{X}) = \sum_{n=1}^{\mathrm{N}} \log \left[\sum_{j=1}^{\mathrm{J}} \pi_j \prod_{i=1}^{\mathrm{d}} \theta_{ji}^{x_{ni}} (1 - \theta_{ji})^{1-x_{ni}} \right]. \tag{3.3}$$

The EM algorithm can be used to learn the maximum likelihood parameters of mixture model of Bernoulli distributions by maximising the likelihood in the Equation (3.3) [167].

3.2 Expectation Maximisation Algorithm

Expectation Maximisation (EM) algorithm is an iterative algorithm to determine the maximum likelihood (MLE) or maximum a posteriori (MAP) estimates of the parameters of latent variable models [39, 114]. The EM algorithm is a popular algorithm for learning model parameters in probabilistic latent variable models by maximising the marginal likelihood. The iterations of EM algorithm alternate between Expectation step (E–Step) and Maximisation Step (M–Step).

E–step estimates the posterior probability of each component for every data point. Whereas, M–step updates the model parameters for next iteration. Iterations between E and M step produce a succession of monotonically increasing sequence of log–likelihood values for the parameters $\tau = 0, 1, 2, 3, \ldots$ regardless of the starting point $\{\pi^{(0)}, \Theta^{(0)}\}$ [114].

3.3 Model Selection in Mixture Models

Model selection is the process of selecting a model of optimal complexity for the given set of (finite,training) data [32, 68]. In the statistics liter-

ature, model selection is the process of selecting a specific model from a plethora of choices [84]. For example, in classification, model selection may refer to choosing a classification algorithm from different classification algorithms such as Naive Bayes, Decision Trees, and Support Vector Machines. The focus in this thesis is modelling of a heterogeneous chromosomal amplification dataset. Mixture models are the model of choice because of their ability to model heterogeneity and their clustering capabilities. The choice of mixture models is also motivated by their ability to learn the structure of the data better than most other methods because each component distributions capture dominant patterns in the data. Furthermore, mixture models are scientifically proven as learning of mixture models involve well studied statistical inference techniques.

In this thesis, model selection refers to the model structure selection or complexity selection which determines the flexibility of the model to fit or explain the data. In other words, model selection in this context refers to choosing an appropriate level of model complexity in the selected class of model, i.e., mixture model. The complexity parameter in mixture model is the number of component distributions in the mixture model. Model selection, therefore, is the selection of number of components in the mixture model [47].

EM algorithm requires apriori knowledge of the number of components in the mixture model to learn the maximum likelihood parameters from the data [114]. However, the number of component distributions are often unknown apriori. Furthermore, one of the major objectives of machine learning and data mining challanges in the real world can often be restricted to determining the number of components in the mixture model. Hence, model selection is essential to learn a mixture model using the EM algorithm.

A mixture model with large number of mixture components produces larger value for the log–likelihood in Equation (3.3). However, a mixture model with large number of mixture components also overfits the data, and generalises poorly on the future unseen data. Additionally, mixture models with large number of components increase complexity in training of mixture models with respect to both time and memory. In contrast, a mixture model with smaller number of mixture components underfits the data, and is unable to adequately represent the underlying data structure. Therefore, model selection aims to optimise this tradeoff between too simple and complex models.

Related work in Model Selection in Mixture Models

A plethora of criteria and methods have been proposed in the literature to determine the optimal number of mixture components in a mixture model [115]. For example, authors in [30], [46], and [131] provide comprehensive review of deterministic, stochastic and resampling criteria for model selection. Deterministic criteria consists of Akaike Information Criterion (AIC) [6], Bayesian Information Criterion (BIC) [142], Minimum Description Length (MDL) [137], and integrated classification likelihood (ICL) [22]. Similarly, stochastic methods includes Markov Chain Monte Carlo (MCMC) [20], and resampling methods includes bootstrapped likelihood ratio test [112]. Similarly, authors in [168] propose a robust approach against model misspecification leading to a better fitting mixture density based on minimum Hellinger distances. In addition, the authors in [31] and [75] use penalised likelihood method for model selection in mixture model.

Data likelihood is often used as the measure of the quality of mixture models [144]. A well trained mixture model with appropriate number of mixture components estimates the underlying data distribution better and produces high likelihood values for the unseen data. In addition, cross–validation have been popular model validation technique in the literature [56, 121, 149]. Hence, in this thesis we use cross–validated log–likelihood as a criteria for model selection.

3.4 Fast Progressive Training of Mixture Models

The EM algorithm is sensitive to initialisation and susceptible to local optima [114, 169]. One solution to avoid local optima is to run the EM algorithm from different random initialisations and select the model with highest likelihood as the global optimum. Similarly, another solution is to take the average of different runs as general performance of the model [153]. Furthermore, the EM algorithm is computationally expensive because of its slow monotonic convergence property [114]. Therefore, multiple restart strategy is popular method in literature where the EM algorithm is run only for a small number of steps, i.e., not until convergence, generating large number of models. Among those models, the model with maximum likelihood can be selected to continue training until convergence [33].

Similarly, different sophisticated algorithms have been proposed to alleviate the problem of local optima in EM algorithm, for example, using splitting and merging of mixture components [86, 154]. In Publication II, we use merging of mixture components as in [154] to train a series of mixture models. The aim is to aid the model selection algorithm to select a model of appropriate complexity, not to avoid local optima. We train multiple models with highest number of component distributions and select the best models among them to start the chain of mixture models by merging the similar mixture components. The training strategy to generate the chain of mixture models resembles backward elimination methodology in feature selection literature [64]. We initially start with large number of mixture components and progressively merge the similar components until the number of components is 1. We use an approximation of Kullback Leibler (KL) divergence as a measure of similarity between the two components in the mixture model.

3.4.1 Kullback Leibler Divergence and Approximation

Kullback Leibler (KL) divergence is a nonsymmetric measure of difference between two probability distributions [96]. The KL divergence between two given probability distributions P and Q on a finite set $\boldsymbol{X}$ is symmetrized by adding the KL divergence from P to Q and Q to P [83].

$$\begin{aligned} \mathcal{D}_{KL}(P||Q) + \mathcal{D}_{KL}(Q||P) &= \sum_i P(i) log \frac{P(i)}{Q(i)} + \sum_i Q(i) log \frac{Q(i)}{P(i)} \\ &= \sum_i \left[\{P(i) - Q(i)\} log \frac{P(i)}{Q(i)} \right], \end{aligned} \tag{3.4}$$

where i indexes all possible combinations of data elements. Extending the KL divergence in Equation (3.4), the KL divergence between two components of a mixture model for data of dimension, d, indexed by k for two component distributions θ and β have been derived in [2] as:

$$\begin{aligned} KL_{\theta\beta} &= \sum_{i=1}^{2^d} \left[\left\{ \prod_{k=1}^{d} \left(\theta_k^{x_{ik}} (1-\theta_k)^{(1-x_{ik})} \right) - \prod_{k=1}^{d} \left(\beta_k^{x_{ik}} (1-\beta_k)^{(1-x_{ik})} \right) \right\} \right. \\ &\quad \left. \cdot \sum_{k=1}^{d} log \frac{\theta_k^{x_{ik}} (1-\theta_k)^{(1-x_{ik})}}{\beta_k^{x_{ik}} (1-\beta_k)^{(1-x_{ik})}} \right], \end{aligned} \tag{3.5}$$

where x_{ik} denotes an element in kth dimensionality of ith sample in the data matrix. The Equation (3.5) is the sum of a large number of elements. If the dimensionality of the data is 5 then we iterate 32 times and when

the dimensionality is 20, we iterate more than a million times (1,048,576). Moreover, the number of comparisons in a mixture model having J components for data of dimensionality d is $2^d J^2$ which is computationally expensive. Therefore, in Publication II, we derive a computationally efficient approximation of the KL divergence as:

$$KL_{\theta\beta} = \sum_{i \in x^*} \left\{ \prod_{k=1}^{d} \left(\theta_k^{x^*_{ik}} (1-\theta_k)^{(1-x^*_{ik})} \right) - \prod_{k=1}^{d} \left(\beta_k^{x^*_{ik}} (1-\beta_k)^{(1-x_{ik})} \right) \right\}, \quad (3.6)$$

where $X^* = \{x^* : x^* \in \overline{\underline{\boldsymbol{X}}}\}$ is a set of all the unique data samples present in the dataset denoted by $\overline{\underline{\boldsymbol{X}}}$. Here, the summation is approximated only with the samples present in the data. Similarly, we remove the fraction containing the log term from Equation (3.5). In Publication II, we are primarily interested in determining the two closest component distributions in a mixture model. We are not necessarily interested in the exact minimum values of KL divergence between two component distributions in a mixture model. These approximations can inaccurately identify two components as most similar to each other while they differ considerably in the full and accurate KL divergence.

The inaccuracies are in the form of selection of two dissimilar components in mixture models to merge. However, in Publication II, we show that our approximation is good estimate of the full KL divergence in terms of matching the two most similar components distributions. Our approximations, as reported in Publication II, is considerably more accurate (twenty five times) than random matching of the components. Moreover, our approximation are 10,000 times faster than full KL divergence for the data dimensionality twenty. Nevertheless, we compensate for any mismatches by retraining the mixture models after merging the mixture components. The aim of the methodology described in Publication II is not to propose any new model selection criteria but to propose an efficient methodology to train a series of mixture models. The models in the series are similar to each other except for the number of mixture components.

3.4.2 Series of Mixture Models

In the algorithm proposed in Publication II, first, we train a large number of mixture models with large number of mixture components (20 in our experiments). Second, we then calculate the approximated KL divergence among all the pairs of mixture components. The two components with minimum approximated KL divergence are then merged as in [154].

The process of merging of mixture components is iterative and continues until the number of components is 1. Mathematically, the merging of the mixing proportions of two candidate component distributions $\pi_{klmin,1}$, and $\pi_{klmin,2}$ to generate a single component distribution π_{merged} can be expressed as:

$$\pi_{merged} = \pi_{klmin,1} + \pi_{klmin,2}. \tag{3.7}$$

Merging the mixture components using Equation (3.7) preserves the properties of mixing proportions such that they have to sum to 1. Similarly, we can merge the parameters of two candidate mixture components $\Theta_{klmin,1}$ and $\Theta_{klmin,2}$ weighted with their mixing components to generate parameters for merged component Θ_{merged} as:

$$\Theta_{merged} = \frac{\pi_{klmin,1} \times \Theta_{klmin,1} + \pi_{klmin,2} \times \Theta_{klmin,2}}{\pi_{klmin,1} + \pi_{klmin,2}}. \tag{3.8}$$

The parameters of merged component distributions in Equation (3.8) also satisfy the properties of probability of a random variable, θ such that $0 \leq \theta \leq 1$. The mixture model obtained after merging the mixture components is retrained before next iteration of merge operation. This progressive training and merging results in a series of mixture models as shown in the Figure 3.1.

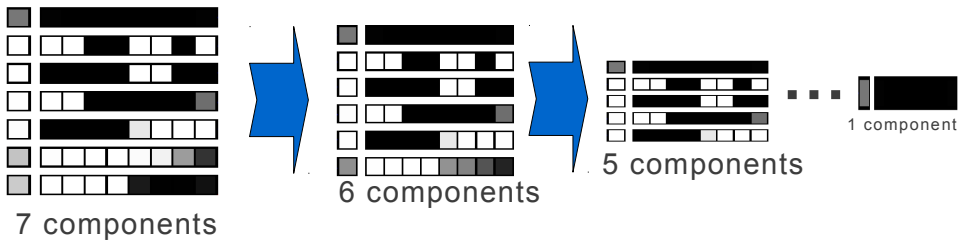

Figure 3.1. Series of mixture models resulting from the progressive merging of the mixture components and retraining of the mixture model. Reprinted with permission from Publication II.

The Figure 3.1 shows snapshot of our algorithm in Publication II where two components in a mixture model with 7 components are merged to generate a mixture model with 6 components. Similarly, mixture models with one less components than the previous model are generated by merging two most similar component distributions until the number of components is 1. The principal focus in Publication II is generating series of mixture models for model selection and not on avoiding local optima or proposing a new model selection criteria.

This series of mixture models can be used with any model selection criteria such as cross–validation, AIC, BIC, and MDL to choose a model of suitable complexity. In our earlier research [2], we have used ten–fold cross–validation to select model of appropriate complexity. We calculate likelihood of each mixture model in the series on both training and validation sets. We then select the model that generalises the best on the validation set taking parsimony into account, i.e., if two models produces comparable results, we select the simpler model [171]. In addition to the gain in computational efficiency, the simple models are also easier to interpret, and understandable to the domain experts [73].

One important property of EM algorithm is that EM algorithm is deterministic for a given initialisation and a given dataset [114]. In other words, if we run EM algorithm on the same data with same initialisation it always converges to the same final model. When the mixture components are merged, the initialisation for the EM algorithm is same. This avoids multiple restarts required in [33] and [153]. Furthermore, EM algorithm converges faster when it is initialised from a merged model than when initialised at random because the merged model resembles the final trained model.

In Publication II, we have shown that EM algorithm converges approximately ten to fifty times faster when initialised from merged model. Similarly, the models produced in the series of models are similar to each other except for the number of components. This allows comparison among similar models for model selection but with different number of components. This avoids the situation when mixture model with some components have been stuck in local minima while models with other components reach global optima. Such situations create a bias in comparison among models with different components in similar vein as ‘unfortunate split’ in cross–validation.

CHAPTER 4

METHODS FOR MULTIRESOLUTION MODELLING

> *With too little data, you won't be able to make any conclusions that you trust. With loads of data you will find relationships that aren't real... Big data isn't about bits, it's about talent.*
>
> — DOUGLAS MERRILL
>
> *Former CIO and VP of Engineering at Google*

Synopsis

The abundance of multiresolution data and increasing benefits of analysing multiple datasets within a single analysis have given major impetus to the research in multiresolution data analysis. In application areas where division of data across different resolutions is smooth, wavelets [81], multiscale methods [11, 163], and scale space theory [100] have been popularly used to analyse multiresolution data. This chapter discusses the core of the thesis and includes most of the scientific contributions of this thesis. This chapter also summarises four of the five publications contained in this thesis.

4.1 Data Transformation

Standard algorithms, such as mixture models, are unable to model multiresolution data in their standard form. Therefore, in Publication I, we propose data transformation methods to analyse multiresolution data by transforming the data to different resolutions and integrating the data in the same resolution. We can then apply the algorithm on the combined data in a single resolution. The methodology of data transformation integrates data in different resolutions and therefore, resembles fusion techniques [28].

Data transformation methods, also called sampling methods, proposed in Publication I are non–stochastic. Sampling resolution in genomics explains the level of precision for measuring the results of a particular experiment: either global (coarse resolution) or detailed (fine resolution). As discussed in Section 2.3 and also shown in the Figure 2.2, the relationship between different resolutions of chromosome, i.e., correspondence of each of the regions in genome in different resolutions are known apriori [143]. We propose two different data transformation methods to transform data across fine and coarse resolution using the knowledge of the correspondence of chromosomal regions in different resolutions.

1. Upsampling transforms the data from coarse resolution to fine resolution increasing the dimensionality of the data. We make multiple copies of a chromosomal region in coarser resolution to upsample the data in coarse resolution to fine resolution.

2. Downsampling transforms the data resolution from fine resolution to coarse resolution decreasing the dimensionality of the data. We downsample using three different methods: OR–function, Weighted, and Majority Decision. We consider the chromosomal amplification pattern of neighbouring chromosomal regions if the number of aberrated chromosomal regions and the number of unaberrated chromosomal regions are equal.

   (a) In OR–function downsampling, a chromosomal region in the coarse resolution is aberrated if any of the chromosomal regions in the fine resolution is aberrated.

   (b) Division of the regions of a chromosome are highly irregular and the

length of a region often differs from the next [143]. In weighted downsampling, a chromosomal region in coarse resolution is aberrated if the total length of the aberrated chromosomal regions is greater than total length of unaberrated chromosomal regions in fine resolution.

(c) In majority decision downsampling, a chromosomal region in the coarse resolution is aberrated if majority of the chromosomal regions in the fine resolution are aberrated.

Experiments on Data Transformation

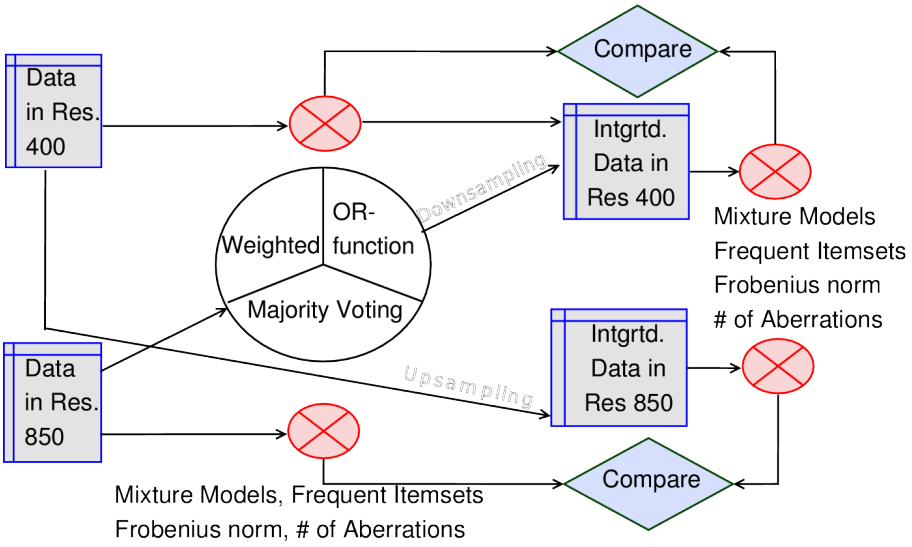

Figure 4.1. Schematic representation of experimental procedure of data transformation methods for multiresolution modelling. First, the data in two different resolutions are transformed to other resolutions. After transformation datasets in the same resolution are integrated. Finally, the algorithm is applied on the integrated dataset. For comparative purposes the algorithm is also applied on the original data before data transformation.

Figure 4.1 depicts the overall experimental procedure where one of the three different downsampling methods transforms the data in fine resolution to coarse resolution. Similarly, a deterministic upsampling method transforms the data in the coarse resolution to the fine resolution. Before data transformation, algorithms such as mixture models, and pattern mining are applied on the data in original resolution. We then integrate the data obtained in same resolution after data transformation. The algorithm is again applied on the integrated data. Finally, we compare the results of the analysis before transformation and after integration in terms of the patterns obtained and model fitting.

Experiments with mixture modelling in different resolutions reported in Publication I show that validation likelihood of the mixture models is higher in the coarser resolution compared to the finer resolution. However, the model selection results are similar across different resolutions as similar number of components are selected in both the coarse and the fine resolution. Although similar number of components are selected, mixture models in coarse resolution produces better likelihood values than the data in fine resolution. In addition, time complexity is higher in the models in the finer resolution. This degradation of performance in fine resolution data can be attributed to the "curse of dimensionality" phenomenon [17], or Hughes effect [76]. Models in coarse resolution are also suitable for understanding and interpreting the results [73].

The results in Publication I also show that the mixture models produce better results on the combined data with the similar number of components than the standalone data in single resolution. This proves the property of mixture model which states that number of components in the mixture model scales with the complexity of the data not with the sample size of the data [65]. The increased sample size arising from the integration of data from other resolution helps nullify the curse of dimensionality and constrains the complexity of mixture models, and avoid overfitting.

MAFIA (MAximal Frequent Itemset Algorithm) [26] was used to mine maximal frequent itemsets in data in the original resolution and the sampled resolution to determine if the data transformation methods preserves the patterns in the data. The results in Publication I show that data transformation across resolutions preserves the maximal frequent itemsets with negligible differences. The negligible differences are expected because data in coarse resolution cannot subsume all the information of the data in fine resolution.

In our earlier research [1], we have also tested the statistical significance of the frequent itemsets (not the maximal frequent itemset) to show that data transformation across different resolution preserves the statistically significant patterns present in the data. In addition, results in [1] also show that statistically significant patterns are also preserved by the generative property of mixture models in all the resolutions. We also compare three different downsampling methods using metrics such as the Frobenius norm [148]. Experimental results in Publication I show that the resulting data produced by three downsampling methods are similar to each other; the variation, if any are negligible.

4.2 Merging of Mixture Components

In Publication III, we use merging of the mixture components of different mixture models in different resolutions to model the multiresolution data. Mixture models can also be used in clustering where component distributions are used as clusters in the data. Proposed multiresolution modelling algorithm resembles clustering aggregation algorithm in [59]. The similarity with clustering aggregation is that we use multiple clustering results, i.e., mixture models to improve the mixture modelling. However, clustering aggregation clusters single dataset generating results as a single clustering solution. In contrast, the proposed multiresolution modelling algorithm analyses different datasets in different resolutions generating clustering solutions in different resolutions.

In the proposed multiresolution modelling algorithm, we first apply mixture models on the data in each resolution separately. Secondly, we iteratively merge the similar mixture components in different mixture models in different resolutions. This is unlike Publication II where we merge the components from the same mixture model. We extend the derivation of fast approximation of Kullback Leibler divergence as a criterion in Publication II to determine the similarity between the mixture components to two mixture models as:

$$\begin{aligned} KL &= \sum_{i \in X^*} \pi_\alpha \prod_{m=1}^{d} \left(\alpha_m^{X^*_{im}} (1-\alpha_m)^{(1-X^*_{im})} \right) \\ &\quad - \sum_{i\prime \in Y^*} \pi_\beta \prod_{n=1}^{d'} \left(\beta_n^{Y^*_{i\prime n}} (1-\beta_n)^{(1-Y^*_{i\prime n})} \right). \end{aligned} \tag{4.1}$$

The approximation of KL divergence in Equation (4.1) resembles Equation (3.6) but for the two component distributions α and β which are components of two different mixture models in different resolutions. Similarly, X^* and Y^* are the unique samples of data in two different resolutions.

We calculate the pairwise KL divergence between all the components in two mixture models. We then select the similar components using minimum weight bipartite matching algorithm [164] as shown in Figure 4.2. The similar components are merged preserving the properties of component distributions and probability of random values in the mixture model. We iterate the matching shown in Figure 4.2 and merging of mixture components until the KL divergence is small enough. Finally, we retrain the

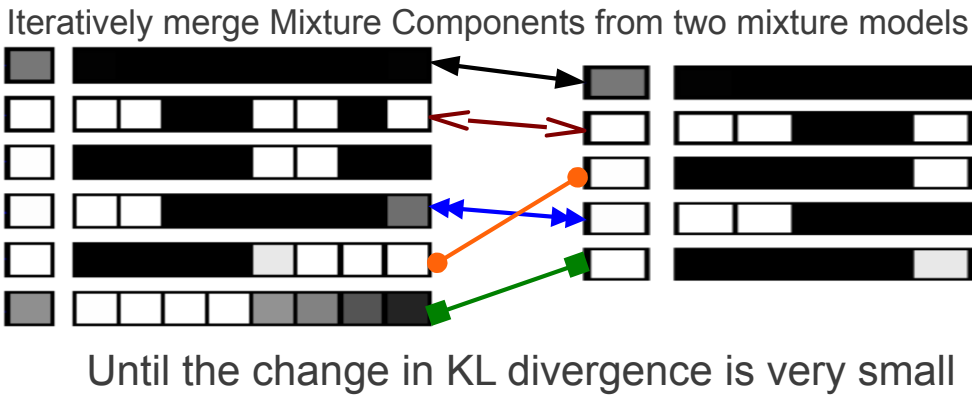

Figure 4.2. Simplified picture of multiresolution modelling using merging of mixture components. We iteratively merge the similar components from different models until the change in KL divergence is very small. The different arrow shapes show the pairwise similarity of mixture components.

mixture models in each resolution. Although mixture models are generated separately in each resolution, they incorporate information about the data in other resolutions.

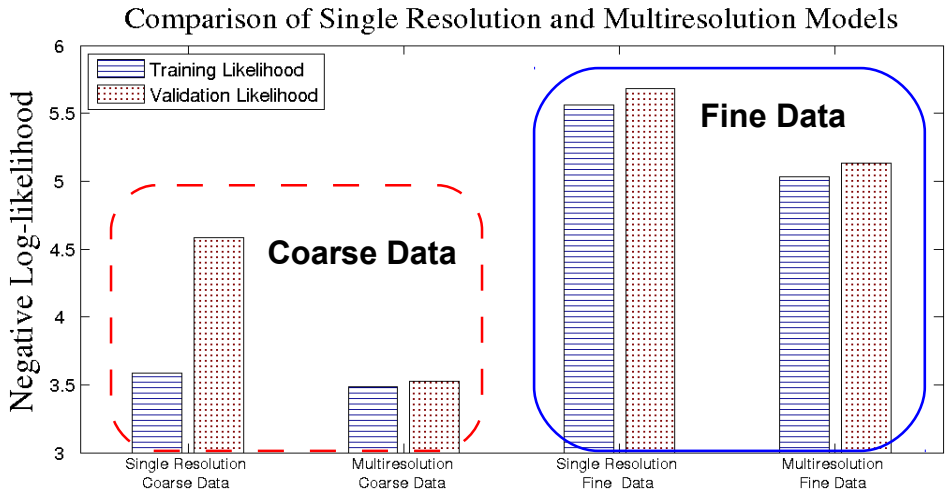

Figure 4.3. Likelihood of multiresolution mixture models trained by merging of mixture components and individually trained mixture models in single resolution. Since the units in Y–axis is the negative log–likelihood, the shorter the bar the better the result. The improvement gained by multiresolution mixture model in the fine resolution is greater than that gained in the coarse resolution. The figure is adapted from Publication III.

The algorithm generates plausible results when the algorithm is experimented on multiresolution chromosomal amplification datasets discussed in Section 2.3. The bar diagram in the Figure 4.3 depicts the improvements gained by multiresolution models over single resolution models. The figure shows training and validation likelihood of the multiresolu-

tion and independent single resolution mixture models trained in a 10–fold cross–validation setting. Since the units in Y–axis is negative log–likelihood, the shorter the bar better the result. The Figure 4.3 shows the two different conditions of the likelihood: first, the performance of single resolution, and the multiresolution model on the coarse data, which is enclosed in the dashed rectangle in the left side of the Figure 4.3. Second, Figure 4.3 also shows performance of the single resolution and the multiresolution models on the fine data which is enclosed in solid rectangle in the right side of the figure.

Scrutinising the results inside both the dashed and the solid rectangles in the Figure 4.3, the performance of the multiresolution model is markedly better in the coarse resolution and slightly better in the fine resolution. The improvement in the performance of the multiresolution model in the coarse resolution is greater than that in the fine resolution. This is because the number of data samples is comparatively smaller in the coarse resolution to add more information to the model in the fine resolution. The results also show that the models trained in the multiresolution setting generalises better on the future unseen data. As discussed in Section 4.2, the performance of the models are better in coarse resolution because of the curse of dimensionality. We also performed the two–tailed t–test to ascertain the statistical significance of the result on the data likelihood [160]. The results also show that both the validation and the training likelihoods are statistically significant when the significance level, α, is 0.1.

4.3 Multiresolution Mixture Components from Domain Ontology

Multiresolution data often forms hierarchical structure as discussed in Chapter 2. The domain ontology used in this thesis is known apriori from the application area. Consequently, we can exploit this structural information from the application area to determine the relationships between data resolutions. Therefore, we can determine the structure of the Bayesian network as shown in the Figure 4.4 with some realistic assumptions for computational efficiency. For this reason, we do not learn the structure of Bayesian networks in our contribution. The assumptions are that the data features in the coarse resolution form the root and branches near the root of the Bayesian network. Similarly, the data features in the finer resolutions form the branches towards the leaves and the leaves

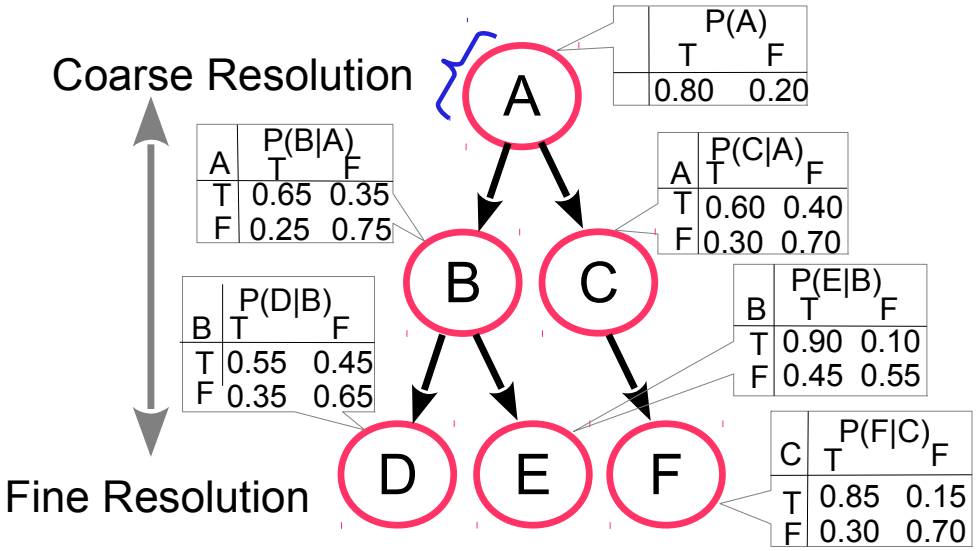

Figure 4.4. A structure of the Bayesian network from the apriori domain knowledge shown in Figure 1.1. The figure shows both Bayesian Network with nodes and edges; and the associated conditional probability tables. The figure is adapted from Publication IV.

of the Bayesian network. Additionally, we can assume that the directed arrows originate from the features in the coarse resolution.

Figure 4.4 shows a Bayesian network generated from the hierarchical structure of data depicted in the Figure 1.1. In the real world, although the hierarchical structure as shown in the Figure 1.1 are known, data in all the resolutions in the structure may not be available. Nevertheless, Bayesian networks have been known for their prowess in missing value imputation [40]. Therefore, in Publication IV, Bayesian networks in the component distributions are used to impute missing data resolutions. Experimental results in Publication IV show that Bayesian networks satisfactorily imputes missing data resolutions.

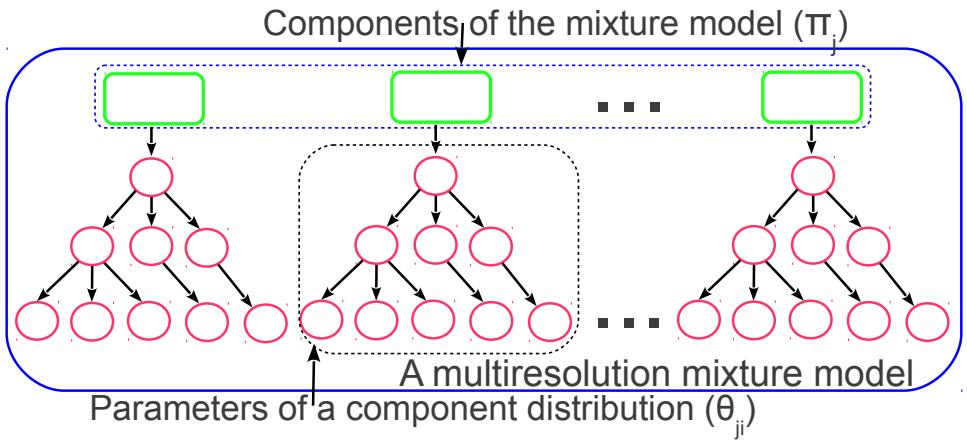

Figure 4.5. Structure of multiresolution mixture model whose components are Bayesian networks. The figure is adapted from Publication IV.

Figure 4.5 depicts the structure of the proposed mixture model in Publication IV for data in multiple resolutions. The three solid rectangles on the top represent different mixture coefficients, π. Similarly, the three network of nodes denote the three component distributions. Each node defines a parameter of the component distributions, θ. The structure of the component distribution is determined from the domain knowledge. Since, the structure of Bayesian network is known, the parameters of these Bayesian networks can be learned in the maximum likelihood framework [9]. If some of the data are missing, we need some assumptions to learn the parameters of the Bayesian networks. One of such similar assumption is founded from the Potts model [15, 135] where we estimate the CPD of the child (C) given parent (P) as: $P(C \mid P) = 0.9$.

After learning the Bayesian networks, and imputing the missing values, the next step is to learn the mixture models. First of the challenges confronting the learning of mixture model is the model selection, i.e., determining the optimal number of component distributions [47]. Similarly, learning the parameters of the component distributions involves learning the parameters of those networks. In general framework for the EM algorithm, we can assign only a single probability value to a node in the mixture model [39]. However, each variable in Bayesian network consists of minimum of two probability values denoting the CPD of the nodes. Hence, we learn the mixture model in the two step procedure. First, we learn the the parameters of individual Bayesian networks in the framework of Bayesian networks [9, 70]. Second, we transform the networks to vectors to learn the parameters of mixture model using the EM algorithm as in [153].

In addition to the multiresolution chromosomal amplification datasets discussed in Section 2.3, we have in Publication IV experimented with a simulated dataset that allows observation of complete data without missing resolutions. The bar diagram in the Figure 4.6 displays the performance of the multiresolution mixture model trained in a 10–fold cross–validation setting and also three different single resolution mixture models trained individually in each resolution. Since units used in the Y–axis is negative log–likelihood, the shorter the bar, better the result. The Figure 4.6 shows two different conditions of likelihood: training and validation. However, the results do not depict change in training and validation likelihood during model selection instead they show the difference in training and validation likelihoods after the selection of components.

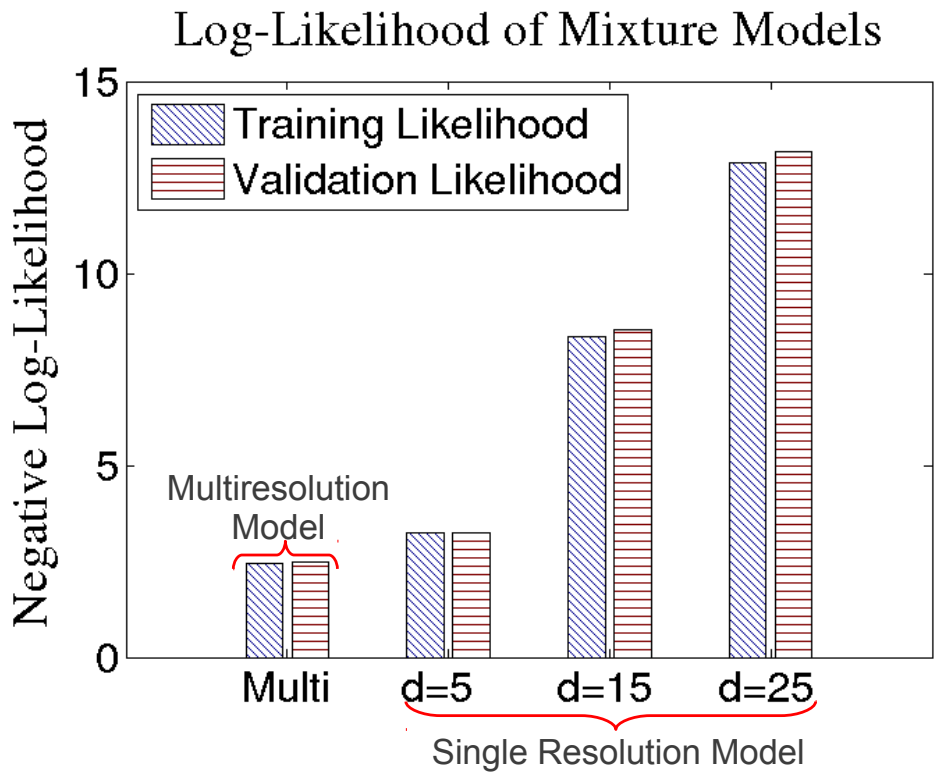

Figure 4.6. Likelihood of single resolution and multiresolution mixture model on simulated dataset. Since the units in Y–axis is the negative log–likelihood, shorter the bar better the result. The performance of multiresolution mixture models surpasses that of all the single resolution models. Reprinted with permission from Publication IV.

The Figure 4.6 shows that the performance of the multiresolution mixture model is markedly better than the three single resolution models. Log–likelihood is comparatively poor in dimensionality of 15, and 25 because of the larger data dimensionality demonstrating curse of dimensionality. The likelihood of the proposed multiresolution model is better than the data with the smallest dimensionality of five in single resolution. The results show that proposed multiresolution mixture model produces plausible results in addition to providing single analysis solution for the data in multiple resolutions.

4.4 Multiresolution Semantic Subgroup Discovery

As discussed in Section 2.4, semantic data mining methods have been gaining popularity in the data mining domain. Similarly, banded matrices have also found usage in data mining domain [43, 55]. In Publication V, we comprehensively analyse multiresolution data using a three stage methodology depicted in Figure 4.7. In the contribution, we ex-

plain the clustering generated by mixture models using semantic data mining methods, and visualise the clusters and the semantic rules using the banded matrices.

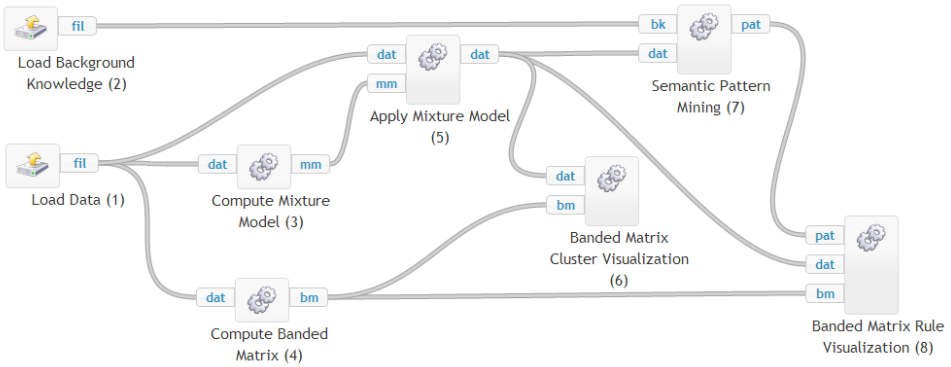

Figure 4.7. The workflow for comprehensive analysis of multiresolution data using a combination of probabilistic model based clustering, semantic data mining, and banded matrices. Reprinted with permission from Publication V.

Figure 4.7 depicts the working of the three part methodology. The figure shows that input to the methodology is the empirical data and additional background knowledge. The additional background knowledge is used by the semantic data mining algorithm to supplement the analysis of the empirical data. As cancer is a heterogeneous and multifactorial disease [90], we use additional background knowledge with an aim to better understand and interpret the results. The additional knowledge provided to the semantic data mining algorithm comprises of fragile sites [42, 141], cancer genes [49], amplification hotspots [125], and virus integration sites [88, 172]. Finally, the taxonomies of hierarchical regions of chromosomes discussed in Section 2.3 are also used as additional background knowledge so that semantic data mining methods are able to analyse multiresolution data.

Mixture models provide an ability to cluster the data considering the components in the mixture model as a cluster [113, 118]. We train the mixture model in a ten–fold cross–validation setting taking parsimony into account [153]. The results produced by mixture models are complex to explain to the application area specialist. Efforts have, however, been made in the past to make the results understandable to the domain experts [73]. In Publication V, we explain the clusters with the rules generated by semantic data mining algorithms and visualisation produced by banded matrices. The cluster labels generated using clustering from mixture model are used as class labels in semantic pattern mining algorithm along with the additional background knowledge. We use general purpose

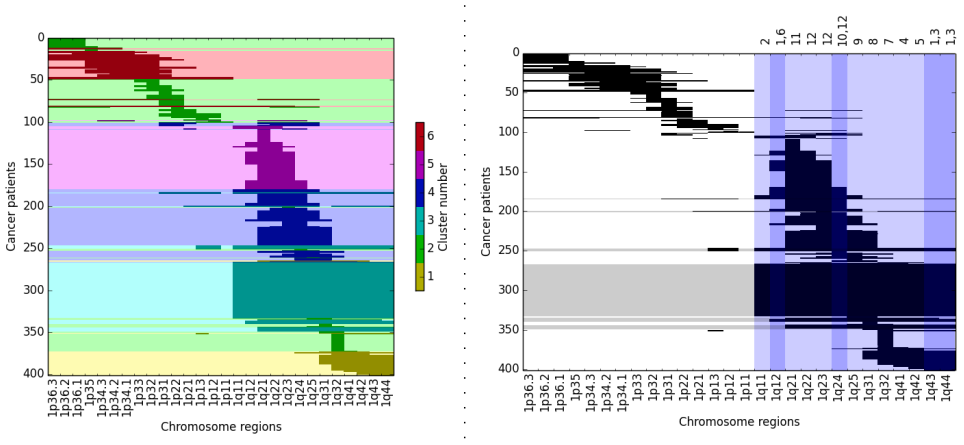

Figure 4.8. The comprehensive analysis of multiresolution data using a combination of probabilistic model based clustering, semantic pattern mining, and banded matrices. Figure on the left panel depicts the clusters overlayed on the banded structure of data. Similarly, figure on the right panel depicts the both clusters and semantic rules overlayed on the banded structure of the data. For the clarity of presentation, the figure on the right depicts only cluster 3 and the rules explaining only cluster 3. Figures are adapted from Publication V.

semantic subgroup discovery system, Hedwig, to find a hypothesis (a predictive model or a set of descriptive patterns) in domain ontology terms, given the training data and the domain knowledge in the form of ontologies [157]. Hedwig, for instance, is developed by the collaborators in Jožef Stefan Institute in Slovenia who are the co–authors in Publication V.

We use the constrained banded matrices [55] to visualise the data. In chromosomal amplification data, the matrices are constrained because the columns denote the specific and unchangeable chromosome regions. We, therefore, shuffle only the rows, i.e., only the samples and not the columns. We then overlay the cluster information along the rows of the banded matrix as shown in the left panel of the Figure 4.8 showing clear distinction among different clusters. In addition, we overlay the rules generated using the semantic subgroup discovery method as shown in the right panel of the Figure 4.8. The visualised rules are tabulated in Table 4.1. The numbers above the rules on top right corner denote the position of rules in the table. A darker hue means that specific region in chromosome appears in more than one rule denoted by more than one position of the rules in the table. Overlaying all the clusters and the rules for each of the clusters will clutter multitude of information on a single figure compromising the understandability of the visualisation. Therefore, we first visualise all the clusters in the data overlaying it on a banded matrix as shown in the left panel of the Figure 4.8. Second, we visualise only a single cluster

#	Rules for cluster 3	TP	FP	Precision
1	Cluster3(X) ← 1q43-44(X) ∧ 1q12(X)	81	0	1.00
2	Cluster3(X) ← 1q11(X)	78	9	0.90
3	Cluster3(X) ← 1q43-44(X)	88	26	0.77
4	Cluster3(X) ← 1q41(X)	88	28	0.76
5	Cluster3(X) ← 1q12(X)	81	43	0.65
6	Cluster3(X) ← 1q32(X)	88	52	0.63
7	Cluster3(X) ← 1q31(X)	87	54	0.62
8	Cluster3(X) ← 1q25(X)	88	64	0.58
9	Cluster3(X) ← 1q24(X)	88	97	0.48
10	Cluster3(X) ← 1q21(X)	88	134	0.40
11	Cluster3(X) ← 1q22-24(X)	88	149	0.37
12	Cluster3(X) ← HotspotSite(X)	88	222	0.28
13	Cluster3(X) ← CancerSite(X)	88	245	0.26
14	Cluster3(X) ← FragileSite(X)	88	259	0.25

Table 4.1. Rules induced for 3 using semantic data mining algorithm Hedwig.

and the rules describing that cluster as shown in the right panel of the Figure 4.8.

The left panel of the Figure 4.8 distinctly shows different clusters proving the credibility of the clustering results. Similarly, the rules visualised in the right panel of the Figure 4.8 identify the amplifications in chromosomal regions that are responsible for certain cluster (cluster 3) and consequently, specific groups of cancer as reported in [126]. In addition, the rules generated by semantic data mining algorithm provide additional insights into the clustering solutions. For example, from the left panel of the Figure 4.8 cluster 3 is denoted by the pronounced amplification in regions 1q11-q44. The rule: Rule 1: Cluster3(X) ← 1q43-44(X) ∧ 1q12(X) characterises 81 out of 88 data samples that are in cluster 3 showing that amplifications in regions 1q43–44 and 1q12 characterises cluster 3 and related cancers with good coverage and precision. Results show that whole region of 1q11–44 need not be aberrated to discriminate that specific cluster of cancers. This provides insights into the data and improvements in the understandability of the amplification to the domain experts.

CHAPTER 5

DISCUSSION

> *Learning is not attained by chance, it must be sought for with ardor and attended to with diligence.*
>
> — ABIGAIL ADAMS
>
> *Letter to John Quincy Adams (1780)*

Synopsis

The work in the thesis focused on the analysis of multiresolution 0–1 data. The application area of choice was chromosomal aberrations patterns in cancer genomics defined in multiple resolutions. The proposed algorithms, mixture models and semantic data mining, for analysis of multiresolution data are experimented on the chromosomal aberrations data with plausible results. Furthermore, an efficient method to train a chain of mixture models was proposed to aid model selection in mixture models. Multiresolution modelling methods, and model selection in mixture models are discussed in this chapter along with their applicability, limitations, and possible future directions of work. The future directions of work discussed in this chapter concerns specific methods discussed and developed in the thesis. The future work section in Chapter 6, i.e. Section 6.2, discusses the overall future work in the multiresolution modelling domain.

5.1 Model Selection in Mixture Models

Model selection is an age–old problem in statistics and machine learning [7, 97]. In Publication II, we do not propose a new model selection

criteria but a computationally efficient method to train a series mixture models differing only in the number of components. The proposed method provides additional facilities of computational efficiency, and similarity of the mixture models in the chain except for the number of components. Therefore, the method is suitable for comparison in model selection. The experiments performed on the three datasets provide evidence of its efficiency and suitability in model selection. Furthermore, the proposed mixture model for Bernoulli distributions can be seamlessly extended to other distributions such as the Gaussian distribution.

The proposed method is sensitive to local optima while learning mixture model via EM algorithm [114, 169]. We try to address the challenges of local optima by training multiple mixture models once for largest number of mixture components before merging the similar components. The best mixture model among the trained mixture models is selected to calculate the KL divergence among mixture components. The most similar components, i.e., the the pair of components with the minimum KL divergence are progressively merged to generate a chain of mixture models. However, this does not guarantee that the EM algorithm reaches global optimum. Avoidance of local minima is still an open research problem in optimisation and also the EM algorithm. Nevertheless, effectiveness, efficiency, and seamless scalability of the proposed method makes the proposed method, the method of choice for training mixture models for model selection.

5.2 Multiresolution Analysis and Modelling of 0–1 Data

Algorithms and methods to study and analyze multiresolution data forms the crux of the thesis. The proposed algorithms complement each other and specific algorithm fulfills the requirements of a specific application. Nevertheless, ample possibilities and challenges for future improvements identified in the proposed algorithms and methods are discussed in the subsequent paragraphs.

5.2.1 Data Transformation for Multiresolution Analysis

The data transformation methods deterministically transform the data across different resolutions in such a way that data in different resolutions can be integrated in a single resolution. The integrated data in

single resolution can then be analyzed using a method of choice because the data is of the same dimensionality. In Publication I, we experiment with mixture models and pattern mining algorithms generating credible results for multiresolution chromosomal aberrations data.

The data transformation methods are suitable for analysis requiring high processing speed and robustness. One of such application area is stream data mining [3, 50, 52, 159] where the requirements are efficient processing and robustness in analysis against minor changes occurring in the data. Data transformation methods are efficient because their computation is simple and are robust against small changes and outliers; for the data transformation methods are deterministic given the structure of the multiresolution phenomena. Furthermore, data transformation methods are suitable for applications requiring single resolution models for multiresolution data. In hindsight, the data transformation methods lack probabilistic interpretation. Adding stochasticity in those methods is a possible future work, for example, with foundations on Potts model [15, 135].

5.2.2 Merging of Mixture Components

In Publication III, we model multiresolution data by generating mixture models in each resolution separately in such a way that the models in each resolution incorporate the information from other data resolutions. The experiments with chromosomal aberrations data show multiresolution mixture models incorporating the interactions between data resolutions produce better results compared to the individually trained single resolution models. The method is suitable for application areas that require models in each level of processing resolution such as image processing, and computer vision [145]. Furthermore, experiments in Publication III have shown that merging of mixture components also helps in avoiding local optima when experimented on the two single resolution models.

Merging of mixture components from different mixture models aids in modelling interaction among the mixture models in different resolutions. An approximation of symmetric KL divergence is used to compare the similarity of the components in the mixture model. The similar components are then merged. However, the convergence analysis of KL divergence is not studied in detail in Publication III. Furthermore, upsampling and downsampling of the parameters of the mixture model adds another complexity to the methodology. Additionally, improvements of the mul-

tiresolution mixture model and avoidance of local optima in single resolution mixture model have been verified only by the empirical experiments. However, solid mathematical foundations and the proofs for the improvement are missing. One direction of future work could focus on mathematical proofs for the empirical evidence in merging components for multiresolution modelling.

5.2.3 Multiresolution Mixture Components

In Publication IV, a single multiresolution mixture model with multiresolution mixture components are proposed and experimented using multiresolution chromosomal aberrations dataset. Only a single multiresolution model is generated in Publication IV, which is unlike Publication III where a model is generated for each data resolution. The individual mixture components provide the functionality of Bayesian networks. The proposed model is suitable for the situations requiring generative modelling prowess of probabilistic models. In Publication IV generative property of the Bayesian network helps imputing the missing resolutions of the data. Furthermore, the proposed multiresolution mixture model could be applicable in any domain where the network structure in the multiresolution data is consistent across the dimensionality, for example, in the image processing domain.

The mixture components used as a Bayesian network model the dependency among the nodes in the network. In addition each node requires at least two probability values describing the probability of the node given the probability of its parent node [9]. In contrast, the EM algorithm assumes IID distributions for the samples [114]. Additionally, the EM algorithm provides only a single probability value for a node, i.e., probability of a random variable (θ) taking the value 1; two if you consider 1 - the given probability ($1 - \theta$). Hence, we transform the nodes to a vector representation to learn the mixture models via the EM algorithm. For this reason, future work in multiresolution mixture modelling could be to develop EM algorithm to directly learn the parameters of mixture models when the components are not vectors but a network. Furthermore, transformation network representation in multiresolution mixture model to vectors and then learning the mixture models using the EM algorithm requires structural similarity of networks used as the different mixture components. Therefore, the future work could focus on relaxing this requirement.

5.2.4 Multiresolution Analysis by Semantic Data Mining

In Publication V, we propose a three part methodology for comprehensive analysis of the multiresolution data. We use clustering results from the mixture models as the labels for the semantic data mining algorithm. The additional background knowledge consists of taxonomy of hierarchy of regions, fragile sites, virus integration sites, amplification hotspots, and cancer genes. We use banded matrices to visualize the clusters from mixture models and the rules from semantic data mining algorithm. The proposed method is suitable for both labeled and unlabeled data as cluster indices can be used as class labels in semantic data mining. Furthermore, banded matrix provides the visualization aspect to the analysis for detailed study of the data. Thus, the method is also suitable for rigorous analysis of multiresolution data.

Every system in the world is connected with one another and each system effects the other system. Consequently, understanding one system can help understand another system better. In this scenario, knowledge or understanding of one system can be used as a background information to understand another system. These methods are applicable in bioinformatics as interacting systems produce different datasets. Similarly, the proposed methodology could be applicable in natural language processing [107] because the additional background knowledge in natural language processing are available in form of ontologies such as the semantic web.

The three part methodology proposed in Publication V takes as an input only data in a single resolution. Multiresolution analysis is achieved by using the taxonomy of multiresolution hierarchy as an additional background knowledge to the methodology. In the future, the semantic pattern mining algorithms can be developed to include data in multiple resolutions simultaneously in addition to the taxonomy of hierarchy of regions.

CHAPTER 6

SUMMARY AND CONCLUSIONS

> *A story has no beginning or end: arbitrarily one chooses that moment of experience from which to look back or from which to look ahead.*
>
> — GRAHAM GREENE
> *The End of the Affair (1951)*

Synopsis

This chapter summarizes the contributions of the thesis and draws conclusion from the research. The chapter also discusses the overall future research perspectives in multiresolution modelling domain.

6.1 Summary

In traditional machine learning and data mining scenario data analysed is from a single source represented in a single resolution. In current age of big data, the challenge is to analyse massive set of datasets, i.e., the challenge is to analyse multiple datasets within a single analysis. The multiple datasets can be available in different representations. Analysis of data in multiple representations needs methods and algorithms suitable for different situations and application areas. Analysis of data in multiple representations within a single analysis framework also caters the needs of data hungry algorithms.

The work in this thesis has concentrated in developing algorithms and methods to address the challenges in modelling data in multiple representations. In this thesis, multiple representations aspect is provided by the data represented in multiple resolutions. The algorithms especially covers mixture models and semantic data mining methods. Different methods and algorithms have been developed to analyse multiresolution data suitable for different situations and application areas.

The data transformation methods proposed in the thesis transforms data across different resolutions to integrate datasets in different resolutions providing an opportunity to analyse data in a single resolution. Additionally, a computationally efficient algorithm to train a series of mixture models to aid model selection algorithms is developed in the thesis. Similarly, an algorithm based on merging of mixture components to model multiresolution data produces models in each resolution incorporating information from other data resolutions. In addition, a multiresolution mixture model uses the domain knowledge to design multiresolution mixture components which are individually functional as Bayesian networks. Furthermore, a semantic data mining algorithm developed in this thesis uses knowledge of hierarchy of multiresolution data and other background knowledge to extract rules from the data. The algorithms and methods provide plausible improvements in multiresolution data analysis compared to the individual analysis in the single resolution data.

6.2 Future Work

The multiresolution analysis methodology developed in this thesis are at its initial stage. The thesis forms the foundations for multiresolution modelling and the algorithms and methods proposed in the thesis need further research on the scope and general applicability. The methods are tested only on datasets such as the chromosomal aberrations datasets, publicly available datasets, and simulated datasets. However, the methods have not been developed as a tool with rigorous testing for general applicability. The improvements necessary for each of the developed methods and algorithms are discussed in Chapter 5. This section discusses the future improvements in overall multiresolution analysis domain. It includes developing the EM algorithm to learn the multiresolution components of the mixture models. The EM algorithm used in this thesis learns the maximum likelihood parameters when networks were arranged as vectors.

Throughout the thesis, mixture models are used in hard clustering setting, i.e., one sample is only associated with one component distribution generating the maximum posterior probability. Mixture models can also be used in a soft clustering setting where posterior probability can be used to assign a sample to more than one component distribution. Soft clustering setting is beneficial in the chromosomewise analysis of chromosomal aberrations data because some cancer samples with the same known cancer labels can be grouped in two different clusters. Soft clustering of chromosomal aberrations data can also be justified because of the heterogeneous nature of cancer.

In chromosomewise analysis, two exactly similar cancer samples can be labelled as two different cancers because other chromosomes that are likely to discriminate cancers will be ignored in the current analysis. Furthermore, we have 73 different types of cancer labels for data in coarse resolution. Therefore, we can use multiclass classification to analyse the data. In a broader context, multiresolution multiclass classification can be a way forward in analysis of multiresolution data.

We need to consider multiresolution data because of the large number of cancer types and smaller number of samples making multiclass classification a challenging task. Furthermore, labels are unavailable for data in fine resolution. In such situations, learning from ambiguous labels [77] or partial labels [35] using clustering labels or the cancer types can help in the analysis of chromosomal aberrations data. Finally, analysis of multiresolution modelling also requires visualisation of the data as well as the results. Therefore, visualisation is also another direction for future work. In Publication V, we use banded matrix to visualise rules and cluster only in single resolution. Initial ideas to visualise multiresolution can borrow from a popular visualisation method in information visualisation known as the Fish eye view [48]. Similar to multiresolution data, Fish eye view also visualises data, providing users a detailed and also a global view.

Bibliography

> *If I have seen further, it is by standing on the shoulders of giants.*
>
> — Isaac Newton
> *In a letter to his rival Robert Hooke (1676)*

[1] P. R. Adhikari and J. Hollmén. Preservation of Statistically Significant patterns in Multiresolution 0-1 Data. In T. Dijkstra, E. Tsivtsivadze, E. Marchiori, and T. Heskes, editors, *Proceedings of the 5th IAPR International Conference on Pattern Recgnition in Bioinformatics*, volume 6282 of *Lecture Notes in Computer Science*, pages 86–97, Nijmegen, The Netherlands, September 2010. Springer Berlin / Heidelberg.

[2] P. R. Adhikari and J. Hollmén. Fast Progressive Training of Mixture Models for Model Selection. In J.-G. Ganascia, P. Lenca, and J.-M. Petit, editors, *Proceedings of Fifteenth International Conference on Discovery Science (DS 2012)*, volume 7569 of *Lecture Notes in Artificial Intelligence*, pages 194–208. Springer-Verlag, October 2012.

[3] C. C. Aggarwal. *Data Streams: Models and Algorithms*, volume 31 of *Advances in Database Systems*. Springer, illustrated edition, 2007.

[4] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, New York, NY, USA, 1993. ACM.

[5] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, San Francisco, CA, USA, 12-15 September 1994. Morgan Kaufmann Publishers Inc.

[6] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *2nd International Symposium on Information Theory*, pages 267–281. Akademiai Kiado, 1973.

[7] T. Ando. *Bayesian Model Selection and Statistical Modeling*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis, 2010.

[8] S. Babaei, M. Hulsman, M. J.T. Reinders, and J. de Ridder. Detecting recurrent gene mutation in interaction network context using multi–scale graph diffusion. *BMC Bioinformatics*, 14:29, January 2013.

[9] D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

[10] A. Bargiela and W. Pedrycz. *Granular Computing: An Introduction*. The Springer International Series in Engineering and Computer Science. Springer US, 2003.

[11] T. J. Barth, T. Chan, and R. Haimes. *Multiscale and Multiresolution Methods: Theory and Applications*. Lecture Notes in Computational Science and Engineering. Springer, 2002.

[12] J. M. S. Bartlett and D. Stirling. A Short History of the Polymerase Chain Reaction. In J. M.S. Bartlett and D. Stirling, editors, *PCR Protocols*, volume 226 of *Methods in Molecular Biology*, pages 3–6. Humana Press, 2003.

[13] M. Baudis. Online database and bioinformatics toolbox to support data mining in cancer cytogenetics. *Biotechniques*, 40(3):269–272, March 2006.

[14] M. Baudis. Genomic imbalances in 5918 malignant epithelial tumors: An explorative meta-analysis of chromosomal CGH data. *BMC Cancer*, 7(1):226, December 2007.

[15] R. J. Baxter. *Exactly Solved Models in Statistical Mechanics*. Dover Publications, January 2008.

[16] S. D. Bay and M. J. Pazzani. Detecting Group Differences: Mining Contrast Sets. *Data Mining and Knowledge Discovery*, 5(3):213–246, 2001.

[17] R. E. Bellman. *Adaptive control processes - A guided tour*. Princeton University Press, Princeton, New Jersey, U.S.A., 1961.

[18] D. Bellot and P. Bessiére. Approximate Discrete Probability Distribution Representation using a Multi–Resolution Binary Tree. In *Proceedings of 15th IEEE International Conference on Tools with Artificial Intelligence, 2003*, ICTAI '03, pages 498–503, 2003.

[19] P. A. Benn and M. A. Perle. Chromosome staining and banding techniques. In D. E. Rooney and B. H. Czepulkowski, editors, *Human Genetics: A Practical Approach*, pages 57–84. IRL Press, 1992.

[20] H. Bensmail, G. Celeux, A. E. Raftery, and C. P. Robert. Inference in model-based cluster analysis. *Statistics and Computing*, 7(1):1–10, 1997.

[21] M. Bianchini, M. Maggini, and L. Sarti. Object Recognition Using Multiresolution Trees. In D-Y Yeung, J. T. Kwok, A. Fred, F. Roli, and D. Ridder, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, volume 4109 of *Lecture Notes in Computer Science*, pages 331–339. Springer Berlin Heidelberg, 2006.

[22] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, July 2000.

[23] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, Secaucus, NJ, USA, 2006.

[24] J. F. Bishop. *Cancer facts : a concise oncology text*. Harwood Academic Publishers, Amsterdam, The Netherlands, 1999.

[25] S. Brenner, M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S. R. Williams, K. Moon, T. Burcham, M. Pallas, R. B. DuBridge, J. Kirchner, K. Fearon, J. Mao, and K. Corcoran. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology*, 18(2000):630–634, June 2000.

[26] D. Burdick, M. Calimlim, and J. Gehrke. MAFIA: a maximal frequent itemset algorithm for transactional databases. In D. Georgakopoulos and A. Buchmann, editors, *Proceedings of 17th International Conference on Data Engineering, 2001*, pages 443–452, 2001.

[27] R. Carlson. The pace and proliferation of biological technologies. *Biosecurity and bioterrorism : biodefense strategy, practice, and science*, 1(3):203–214, 2003.

[28] D. B. Carter. *Analysis of multiresolution data fusion techniques*. PhD thesis, Virginia Polytechnic Institute and State University, 1998.

[29] R. Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, July 1997.

[30] G. Celeux. Mixture Models for Classification. In R. Decker and H-J. Lenz, editors, *Advances in Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 3–14. Springer Berlin Heidelberg, 2007.

[31] J. Chen and A. Khalili. Order Selection in Finite Mixture Models With a Nonsmooth Penalty. *Journal of the American Statistical Association*, 103(484):1674–1683, 2008.

[32] V. S. Cherkassky and F. Mulier. *Learning from Data: Concepts, Theory, and Methods*. John Wiley & Sons, Inc., New York, NY, USA, first edition, 1998.

[33] D. M. Chickering and D. Heckerman. Efficient Approximations for the Marginal Likelihood of Bayesian Networks with Hidden Variables. *Machine Learning*, 29(2-3):181–212, November 1997.

[34] S. M. Cohen. Aristotle's Metaphysics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford Online, spring 2014 edition, 2014.

[35] T. Cour, B. Sapp, and B. Taskar. Learning from Partial Labels. *Journal of Machine Learning Research*, 12:1501–1536, July 2011.

[36] K. Crammer, M. Kearns, and J. Wortman. Learning from Multiple Sources. *Journal of Machine Learning Research*, 9:1757–1774, August 2008.

[37] "datum data". *Merriam-Webster's dictionary of English usage*. Merriam-Webster, Springfield, Massachusetts, 2002.

[38] J. de Ridder, J. Kool, A. Uren, J. Bot, L. Wessels, and M. J. T. Reinders. Co–occurrence analysis of insertional mutagenesis data reveals cooperating oncogenes. *Bioinformatics*, 23(13):i133–41, July 2007.

[39] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38, 1977.

[40] M. Di Zio, M. Scanu, L. Coppola, O. Luzi, and A. Ponti. Bayesian networks for imputation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167(2):309–322, May 2004.

[41] U.S. DOE. *New Frontiers in Characterizing Biological Systems: Report from the May 2009 Workshop*. DOE/SC-0121. U.S. Department of Energy, Office of Science, 2009.

[42] S. G. Durkin and T. W. Glover. Chromosome Fragile Sites. *Annual Review of Genetics*, 41(1):169–192, 2007.

[43] L. Eldén. *Matrix Methods in Data Mining and Pattern Recognition*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2007.

[44] Electronics, Electrical Engineering Laboratory (National Institute of Standards, and Technology). *Measurements for competitiveness in electronics [microform] / prepared by the Electronics and Electrical Engineering Laboratory*. The Laboratory ; National Technical Information Service [distributor Gaithersburg, MD : Springfield, VA, first edition, 1993.

[45] B. S. Everitt and D. J. Hand. *Finite mixture distributions*. Chapman and Hall, London; New York, 1981.

[46] M. A. T. Figueiredo and A. K. Jain. Unsupervised Learning of Finite Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, March 2002.

[47] C. Fraley and A. E. Raftery. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal*, 41(8):578–588, 1998.

[48] G. W. Furnas. Generalized Fisheye Views. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '86, pages 16–23, New York, NY, USA, 1986. ACM.

[49] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton. A census of human cancer genes. *Nature reviews. Cancer*, 4(3):177–183, Mar 2004.

[50] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy. Mining Data Streams: A Review. *SIGMOD Record*, 34(2):18–26, June 2005.

[51] A. Gallo, P. Miettinen, and H. Mannila. *Finding Subgroups having Several Descriptions: Algorithms for Redescription Mining*, chapter 29, pages 334–345. SIAM, 2008.

[52] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A Survey on Concept Drift Adaptation. *ACM Computing Survey*, 46(4):44:1–44:37, March 2014.

[53] D. Gamberger and N. Lavrač. Expert-Guided Subgroup Discovery: Methodology and Application. *Journal of Artificial Intelligence Research*, 17(1):501–527, 2002.

[54] M. Garland. Multiresolution Modeling: Survey & Future Opportunities. In *Eurographics '99 – State of the Art Reports*, pages 111–131, 1999.

[55] G. C. Garriga, E. Junttila, and H. Mannila. Banded structure in binary matrices. *Knowledge and Information Systems*, 28(1):197–226, 2011.

[56] S. Geisser. A Predictive Approach to the Random Effect Model. *Biometrika*, 61(1):101–107, 1974.

[57] G. Giancarlo, G. Wagner, and M. Marten Sinderen van. A formal theory of conceptual modeling universals. In *1st Intl. Workshop on Philosophy and Informatics, WSPI 2004*, volume 112 of *CEUR workshop proceedings*. DFKI, 2004.

[58] W. Gilbert and A. Maxam. The Nucleotide Sequence of the lac Operator. *Proceedings of the National Academy of Sciences*, 70(12, Part I):3581–3584, December 1973.

[59] A. Gionis, H. Mannila, and P. Tsaparas. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data*, 1(1):341–352, March 2007.

[60] T. W. Glover, M. F. Arlt, A. M. Casper, and S. G. Durkin. Mechanisms of common fragile site instability. *Human Molecular Genetics*, 14(Supplement 2):R197–R205, 2005.

[61] I. R. Goodman, R. P. Mahler, and H. T. Nguyen. *Mathematics of Data Fusion*. Kluwer Academic Publishers, Norwell, MA, USA, 1997.

[62] P. V. Gopalacharyulu, E. Lindfors, C. Bounsaythip, T. Kivioja, L. Yetukuri, J. Hollmén, and M. Orešič. Data integration and visualization system for enabling conceptual biology. *Bioinformatics*, 21(Suppl.1):i177–i185, 2005.

[63] T. R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199 – 220, 1993.

[64] I. Guyon and A. Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182, March 2003.

[65] R. Hammoud and R. Mohr. Biometrics: Promising frontiers for emerging identification market. Technical Report 3905, Unité de recherche INRIA Rhŏne-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN (France), March 2000.

[66] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, January 2007.

[67] D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. Adaptive Computation and Machine Learning Series. MIT Press, 2001.

[68] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, second edition, February 2009.

[69] T. X. He. *Wavelet Analysis and Multiresolution Methods*. Lecture Notes in Pure and Applied Mathematics. Taylor & Francis, 2000.

[70] D. Heckerman. A Tutorial on Learning With Bayesian Networks. In M. I. Jordan, editor, *Learning in graphical models*, pages 301–354. MIT Press, USA, 1999.

[71] F. Herrera, C. J. Carmona, P. González, and M. Jesus. An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, 29(3):495–525, 2011.

[72] J. Hipp, U. Güntzer, and G. Nakhaeizadeh. Algorithms for Association Rule Mining — a General Survey and Comparison. *SIGKDD Explorations Newsletter*, 2(1):58–64, June 2000.

[73] J. Hollmén and J. Tikka. Compact and understandable descriptions of mixture of Bernoulli distributions. In M.R. Berthold, J. Shawe-Taylor, and N. Lavrač, editors, *Proceedings of the 7th International Symposium on Intelligent Data Analysis (IDA 2007)*, volume 4723 of *Lecture Notes in Computer Science*, pages 1–12, Ljubljana, Slovenia, September 2007. Springer-Verlag.

[74] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. P. Hill, R. Kania, M. Schaeffer, S. S. Pierre, S. Twigger, O. White, and S. Y. Rhee. Big data: The future of biocuration. *Nature*, 455(7209):47–50, 2008.

[75] T. Huang, H. Peng, and K. Zhang. Model Selection for Gaussian Mixture Models. *arXiv preprint arXiv:1301.3558*, 2013.

[76] G. Hughes. On the Mean Accuracy of Statistical Pattern Recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, January 1968.

[77] E. Hüllermeier and J. Beringer. Learning from Ambiguously Labeled Examples. *Intelligent Data Analysis*, 10(5):419–439, September 2006.

[78] I. Huopaniemi, T. Suvitaival, J. Nikkilä, M. Orešič, and S. Kaski. Multivariate multi-way analysis of multi-source data. *Bioinformatics*, 26:i391–i398, 2010.

[79] A. Hussain and A. Visilsoanathan. Multiresolution Semantic Visualization of Network Traffic. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*, ICSC '11, pages 364–367, Washington, DC, USA, 2011. IEEE Computer Society.

[80] A. Iske. *Multiresolution Methods in Scattered Data Modelling*. Springer Berlin Heidelberg, first edition, 2004.

[81] B. Jawerth and W. Sweldens. An Overview of Wavelet Based Multiresolution Analyses. *SIAM Review*, 36(3):377–412, September 1994.

[82] V. Jovanoski and N. Lavrač. Classification Rule Learning with APRIORI–C. In P. Brazdil and A. Jorge, editors, *Progress in Artificial Intelligence*, volume 2258 of *Lecture Notes in Computer Science*, pages 44–51. Springer Berlin Heidelberg, 2001.

[83] B. H. Juang and L. R. Rabiner. A probabilistic distance measure for Hidden Markov Models. *AT&T Technical Journal*, 64(2):391–408, February 1985.

[84] J. B. Kadane and N. A. Lazar. Methods and Criteria for Model Selection. *Journal of the American Statistical Association*, 99(465):279–290, March 2004.

[85] A. Kallioniemi, O. P. Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman, and D. Pinkel. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *SCIENCE*, 258(5083):818–821, October 30 1992.

[86] G. Karčiauskas. Learning of Latent Class Models by Splitting and Merging Components. In P. Lucas, J. A. Gómez, and A. Salmerón, editors, *Advances in Probabilistic Graphical Models*, volume 214 of *Studies in Fuzziness and Soft Computing*, pages 235–251. Springer Berlin Heidelberg, 2007.

[87] E. Kettunen, A. G. Nicholson, B. Nagy, J. K. Seppänen, T. Ollikainen, G. Ladas, V. Kinnula, M. Dusmet, S. Nordling, J. Hollmén, D. Kamel, P. Goldstraw, and S. Knuutila. L1CAM, INP10, P-cadherin, tPA and ITGB4 over-expression in malignant pleural mesotheliomas revealed by combined use of cDNA and tissue microarray. *Carcinogenesis*, 26(1):17–25, September 2005.

[88] J. D. Khoury, N. M. Tannir, M. D. Williams, Y. Chen, H. Yao, J. Zhang, E. J. Thompson, F. Meric-Bernstam, L. J. Medeiros, J. N. Weinstein, et al. The Landscape of DNA Virus Associations Across Human Malignant Cancers Using RNA-seq: An Analysis of 3,775 Cases. *Journal of Virology*, 87(16):8916–8926, 2013.

[89] P. M. Kim. *Understanding Subsystems in Biology through Dimensionality Reduction, Graph Partitioning and Analytical Modeling*. PhD thesis, Massachusetts Institute of Technology, February 2003.

[90] R. A. King, J. I. Rotter, and A. G. Motulsky, editors. *The Genetic Basis of Common Diseases.* Oxford Monographs on Medical Genetics. Oxford University Press, second edition, 2002.

[91] I. R. Kirsch. *The causes and consequences of chromosomal aberrations*. CRC Press, 1993.

[92] H. A. Klein. *The Science of Measurement: A Historical Survey*. Dover Books on Mathematics. Dover Publications, 2012.

[93] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo. Finding Interesting Rules from Large Sets of Discovered Association Rules. In N. R. Adam, B. K. Bhargava, and Y. Yesha, editors, *Proceedings of the Third International Conference on Information and Knowledge Management*, CIKM '94, pages 401–407, New York, NY, USA, 1994. ACM.

[94] S. Knuutila, K. Autio, and Y. Aalto. Online Access to CGH Data of DNA Sequence Copy Number Changes. *American Journal of Pathology*, 157(2):689–689, August 2000.

[95] P. Koikkalainen. Progress with the Tree-Structured Self-Organizing Map. In A. G. Cohn, editor, *In Proceedings on 11th European Conference on Artificial Intelligence (ECAI)*, pages 211–215. European Committee for Artificial Intelligence (ECCAI), John Wiley & Sons, Ltd., 1994.

[96] S. Kullback and R. A. Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, March 1951.

[97] P. Lahiri, editor. *Model Selection*. Institute of Mathematical Statistics Lecture Notes Monograph. Institute of Mathematical Statistics, illustrated edition, 2001.

[98] N. Lavrač, A. Vavpetič, L. Soldatova, I. Trajkovski, and P. Kralj Novak. Using Ontologies in Semantic Data Mining with SEGS and g-SEGS. In

T. Elomaa, J. Hollmén, and H. Mannila, editors, *Proceedings of the International Conference on Discovery Science (DS '11)*, volume 6926 of *Lecture Notes in Computer Science*, pages 165–178. Springer Berlin Heidelberg, 2011.

[99] T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, 21(2):224–270, 1994.

[100] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, Norwell, MA, USA, 1994.

[101] B. Liu, W. Hsu, and Y. Ma. Integrating Classification and Association Rule Mining. In *Proceedings of the 4th international conference on Knowledge Discovery and Data mining (KDD'98)*, pages 80–86. AAAI Press, August 1998.

[102] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law. Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, 2012:1–11, 2012.

[103] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(13):1675–1680, 1996.

[104] C. Lynch. Big data: How do your data grow? *Nature*, 455(7209):28–29, September 2008.

[105] S. G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, July 1989.

[106] H. Mannila, H. Toivonen, and A. I. Verkamo. Efficient algorithms for discovering association rules. In U. M. Fayyad and R. Uthurusamy, editors, *AAAI Workshop on Knowledge Discovery in Databases (KDD-94)*, pages 181–192, Seattle, Washington, 1994. AAAI Press.

[107] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, illustrated edition, 1999.

[108] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute, June 2011.

[109] E. R. Mardis. Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 9(1):387–402, June 2008.

[110] E. R. Mardis. A decade's perspective on DNA sequencing technology. *Nature*, 470(7333):198–203, February 2011.

[111] V. Marx. Biology: The big challenges of big data. *Nature*, 498(7453):255–260, June 2013.

[112] G. J. McLachlan. On Bootstrapping the Likelihood Ratio Test Stastistic for the Number of Components in a Normal Mixture. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3):318–324, 1987.

[113] G. J. McLachlan and K. E. Basford. Mixture models. Inference and applications to clustering. *Statistics: Textbooks and Monographs, New York*, 1, 1988.

[114] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley series in probability and statistics. Wiley, second edition, 2008.

[115] G. J. McLachlan and D. Peel. *Finite mixture models*, volume 299 of *Probability and Statistics – Applied Probability and Statistics Section*. Wiley, 2000.

[116] D. W. Mcshea. Complexity and evolution: What everybody knows. *Biology and Philosophy*, 6(3):303–324, July 1991.

[117] M. Meila and M. I. Jordan. Learning with Mixtures of Trees. *Journal of Machine Learning Research*, 1:1–48, Oct 2000.

[118] V. Melnykov and R. Maitra. Finite Mixture Models and Model–Based Clustering. *Statistics Surveys*, 4(1):80–116, 2010.

[119] P. Milanfar. *Super-resolution imaging*. CRC Press, 2010.

[120] T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[121] F. Monsteller and J. W. Tukey. Data analysis, including statistics. In G. Lindzey and E. Aronson, editors, *Handbook of Social Psychology*, volume 2, chapter 10, pages 80–203. Addison-Wesley, Reading, MA, 1968.

[122] A. Moore. Very Fast EM-based Mixture Model Clustering Using Multiresolution KD–trees. In M. Kearns and D. Cohn, editors, *Advances in Neural Information Processing Systems*, pages 543–549. Morgan Kaufman, April 1999.

[123] G. E. Moore. Cramming More Components onto Integrated Circuits. *Electronics*, 38(8):114–117, April 1965.

[124] D. Mukherjee, Q. M. J. Wu, and T. M. Nguyen. Multiresolution Based Gaussian Mixture Model for Background Suppression. *IEEE Transactions on Image Processing*, 22(12):5022–5035, 2013.

[125] S. Myllykangas, J. Himberg, T. Böhling, B. Nagy, J. Hollmén, and S. Knuutila. DNA copy number amplification profiling of human neoplasms. *Oncogene*, 25(55):7324–7332, November 2006.

[126] S. Myllykangas, J. Tikka, T. Böhling, S. Knuutila, and J. Hollmén. Classification of human cancers based on DNA copy number amplification modeling. *BMC Medical Genomics*, 1(15), May 2008.

[127] The Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, September 2013.

[128] S-K. Ng and G. J. McLachlan. Robust Estimation in Gaussian Mixtures Using Multiresolution Kd-trees. In C. Sun, H. Talbot, S. Ourselin, and T. Adriaansen, editors, *Proceedings of the 7th International Conference on Digital Image Computing: Techniques and Applications*, pages 145–154. CSIRO Publishing, 2003.

[129] P. Novak, N. Lavrač, and G. I. Webb. Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining. *Journal of Machine Learning Research*, 10:377–403, February 2009.

[130] G. Obe and Vijayalaxmi. *Chromosomal alterations: methods, results, and importance in human health*. Springer, 2007.

[131] A. Oliveira-Brochado and F. V. Martins. Assessing the Number of Components in Mixture Models: a Review. FEP Working Papers 194, Universidade do Porto, Faculdade de Economia do Porto, November 2005.

[132] P. Panov. *A Modular Ontology of Data Mining*. Doctoral dissertation, Jožef Stefan International Postgraduate School, July 2012.

[133] G. Piatetsky-Shapiro. Discovery, Analysis, and Presentation of Strong Rules. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*. AAAI/MIT Press, 1991.

[134] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20:207–211, 1998.

[135] R. B. Potts. Some generalized order-disorder transformations. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(01):106–109, 1952.

[136] C. K. Reddy and J-H. Park. Multi-resolution Boosting for Classification and Regression Problems. In T. Theeramunkong, B. Kijsirikul, N. Cercone, and T-B. Ho, editors, *Advances in Knowledge Discovery and Data Mining*, volume 5476 of *Lecture Notes in Computer Science*, pages 196–207. Springer Berlin Heidelberg, 2009.

[137] J. Rissanen. Modeling By Shortest Data Description. *Automatica*, 14(5):465 – 471, 1978.

[138] S. W. Roh, G. C. J. Abell, K-H. Kim, Y-D. Nam, and J-W Bae. Comparing microarrays and next-generation sequencing technologies for microbial ecology research. *Trends in Biotechnology*, 28(6):291–299, 2010.

[139] B. Russell. On the Relations of Universals and Particulars. *Proceedings of the Aristotelian Society*, 12:1–24, 1911.

[140] S. Sagiroglu and D. Sinanc. Big data: A review. In *International Conference on Collaboration Technologies and Systems (CTS), 2013*, pages 42–47, 2013.

[141] M. Schwartz, E. Zlotorynski, and B. Kerem. The molecular basis of common and rare fragile sites. *Cancer Letters*, 232(1):13–26, 2006.

[142] G. Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464, March 1978.

[143] L. G. Shaffer and N. Tommerup. *ISCN 2005: An International System for Human Cytogenetic Nomenclature(2005) Recommendations of the International Standing Committee on Human Cytogenetic Nomenclature*. Karger, 2005.

[144] P. Smyth. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10(1):63–72, 2000.

[145] M. Sonka, V. Hlavac, and R. Boyle. *Image processing, analysis, and machine vision*. Thomson, Toronto, 3 edition, 2008.

[146] P. Stankiewicz and J. R. Lupski. Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, 61(1):437–455, 2010.

[147] B. W. Stewart and C. P. Wild, editors. *World Cancer Report 2014*. International Agency for Research on Cancer (IARC) Nonserial, 2008.

[148] G. W. Stewart. *Matrix Algorithms: Volume 1, Basic Decompositions*. Society for Industrial Mathematics, illustrated edition, 1998.

[149] M. Stone. Cross-validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society*, 36(2):111–147, 1974.

[150] S. Sun. A survey of multi–view machine learning. *Neural Computing and Applications*, 23(7–8):2031–2038, 2013.

[151] B. Swartout, R. Patil, K. Knight, and T. Russ. Towards distributed use of large–scale ontologies. In *Proceedings of AAAI Symposium on Ontological Engineering*, pages 138–148, 1996.

[152] N. Tatti. *Advances in Mining Binary Data: Itemsets as Summaries*. PhD thesis, Helsinki University of Technology, Faculty of Information and Natural Sciences, 2008.

[153] J. Tikka, J. Hollmén, and S. Myllykangas. Mixture Modeling of DNA copy number amplification patterns in cancer. In F. Sandoval, A. Prieto, J. Cabestany, and M. Graña, editors, *Proceedings of the 9th International Work-Conference on Artificial Neural Networks (IWANN 2007)*, volume 4507 of *Lecture Notes in Computer Science*, pages 972–979, San Sebastián, Spain, 2007. Springer-Verlag.

[154] N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton. SMEM Algorithm for Mixture Models. *Neural Computation*, 12(9):2109–2128, 2000.

[155] A. Usvasalo, E. Elonen, U. M. Saarinen-Pihkala, R. Räty, A. Harila-Saari, P. Koistinen, E-R. Savolainen, S. Knuutila, and J. Hollmén. Prognostic classification of patients with acute lymphoblastic leukemia by using copy number profiles identified from array-based comparative genomic hybridization data. *Leukemia Research*, 34(11):1476–1482, November 2010.

[156] A. Vavpetič and N. Lavrač. Semantic Subgroup Discovery Systems and Workflows in theSDM-Toolkit. *The Computer Journal*, 56(3):304–320, 2013.

[157] A. Vavpetič, V. Podpečan, and N. Lavrač. Semantic subgroup explanations. *Journal of Intelligent Information Systems*, 42(2):233–254, 2013.

[158] B. Vogelstein and K. W. Kinzler. *The genetic basis of human cancer*. McGraw-Hill, New York, 2002.

[159] I. Žliobaitė and J. Hollmén. Optimizing regression models for data streams with missing values. *Machine Learning*, page In Press, 2014.

[160] R. Walpole, R. Myers, S. Myers, and K. Ye. *Probability & Statistics for Engineers & Scientists*. Prentice Hall, NJ, USA, nineth edition, 2012.

[161] Y. Wang, J. Hayakawa, F. Long, Q. Yu, A. H. Cho, G. Rondeau, J. Welsh, S. Mittal, I. De Belle, E. Adamson, M. McClelland, and D. Mercola. "promoter array" studies identify cohorts of genes directly regulated by methylation, copy number change, or transcription factor binding in human cancer cells. *Annals of the New York Academy of Sciences*, 1058(1):162–185, November 2005.

[162] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, April 1953.

[163] E. Weinan. *Principles of Multiscale Modeling*. Cambridge University Press, 2011.

[164] D. B. West. *Introduction to graph theory*. Prentice Hall, second (illustrated) edition, 1996.

[165] A. S. Willsky. Multiresolution Markov models for signal and image processing. *Proceedings of the IEEE*, 90(8):1396–1458, August 2002.

[166] R. Wilson. MGMM: multiresolution Gaussian mixture models for computer vision. In *Proceedings of 15th International Conference on Pattern Recognition*, volume 1, pages 212–215. IEEE, September 2000.

[167] J. H. Wolfe. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5(3):329–350, July 1970.

[168] M-J. Woo and T. N. Sriram. Robust Estimation of Mixture Complexity. *Journal of the American Statistical Association*, 101(476):1475–1486, December 2006.

[169] C. F. J. Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95–103, March 1983.

[170] J. T. Yao and Y. Y. Yao. A Granular Computing Approach to Machine Learning. In L. Wang, S. K. Halgamuge, and X. Yao, editors, *Proceedings of the 1st International Conference on Fuzzy Systems and Knowledge Discovery: Computational Intelligence for the E-Age, 2 Volumes, November 18-22, 2002, Orchid Country Club, Singapore*, FSDK'02, pages 732–736, 2002.

[171] A. Zellner, H. A. Keuzenkamp, and M. McAleer, editors. *Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple*. Cambridge University Press, March 2009.

[172] H. zur Hausen. The search for infectious causes of human cancers: Where and why. *Virology*, 392(1):1 – 10, September 2009.

Publication I

Prem Raj Adhikari, Jaakko Hollmén. Patterns from Multiresolution 0–1 data. In *Jilles Vreeken, Nikolaj Tatti, and Bart Goethals, Editors, UP '10, ACM SIGKDD Workshop on Useful Patterns*, Washington DC, ACM, New York, NY, USA, Pages 8–16, July 25, 2010, DOI: 10.1007/s10844-013-0282-3, July 2010.

Patterns from multiresolution 0-1 data

Prem Raj Adhikari and Jaakko Hollmén
Aalto University School of Science and Technology
Department of Information and Computer Science
PO Box 15400, FI-00076 Aalto
prem.adhikari@tkk.fi and jaakko.hollmen@tkk.fi

ABSTRACT

Biological systems are complex systems and often the biological data is available in different resolutions. Computational algorithms are often designed to work with only specific resolution of data. Hence, upsampling or downsampling is necessary before the data can be fed to the algorithm. Moreover, high-resolution data incorporates significant amount of noise thus producing explosion of redundant patterns such as maximal frequent itemset, closed frequent itemset and non-derivable itemset in the data which can be solved by downsampling the data if the information loss is insignificant during sampling. Furthermore, comparing the results of an algorithm on data in different resolution can produce interesting results which aids in determining suitable resolution of data. In addition, experiments in different resolutions can be helpful in determining the appropriate resolution for computational methods. In this paper, three methods of downsampling are proposed, implemented and experiments are performed on different resolutions and the suitability of the proposed methods are validated and the results compared. Mixture models are trained on the data and the results are analyzed and it was seen that the proposed methods produce plausible results showing that the significant patterns in the data are retained in lower resolution. The proposed methods can be extensively used in integration of databases.

Keywords

Binary data, multiple resolutions, Upsampling, Downsampling, Mixture Models

1. INTRODUCTION

This paper proposes and studies three different downsampling methods for genome-wide chromosome bands in amplification data. Sample in this context is a process of defining the level of precision for staining the chromosome bands. For example, chromosome-1 can be defined by 23, 28, 42, 61 and 63 bands in resolution 300, 400, 550, 700 and 850 respectively as defined by International System on Cytogentic Nomenclature(ISCN)[1]. The proposed methods can also be used in downsampling of similar data where the data is encoded in different chromosome bands for each sample. Biological systems are very complex systems. Since these complexities are directly related to the health of humans, different technologies have been developed to study them. Microarray technology, such as CGH (Comparative genomic hybridization)[2] and aCGH (Array comparative genomic hybridization) [3] have given the facilities to study the genomes and the genes in human body. Thus, biological data are available in different resolutions. However, computational algorithms can often handle only specific resolution of the data. So, data needs to be upsampled and downsampled to different resolutions before some algorithms are applied on them. Furthermore, comparing the results of the models in different resolutions can reveal some interesting facts useful for cancer research. If data in lower resolution produces results comparable to data in higher resolution, the time, computational and hardware costs required to obtain the data in higher resolution can be saved. Here, a dataset in resolution 850 was downsampled in four different resolutions 300, 393, 550, and 700 and experiments were performed on the data in different resolutions. In addition, a different dataset in resolution 393 was upsampled to resolution 550, 700 and 850 and downsampled to resolution 300. Other popular dimensionality reduction methods[4] does not produce desirable results because representation of the data is lost. In this context, we present methods of upsampling and downsampling of data and experimental results in integrating two biological databases originally in different resolutions.

The major aim of the paper is to sample the data in different resolutions such that the significant patterns i.e. frequent itemsets are retained in the sampled data. Thus, we experiment our proposed methods with a set of pattern mining algorithms. Given a binary data, $\mathcal{D}$ with a set of attributes $\mathcal{I}_1, \mathcal{I}_2 \ldots \mathcal{I}_n$ and a support σ, frequent set is the set $\mathcal{F}$ of items of $\mathcal{D}$ such that at least a fraction of σ of the rows of $\mathcal{D}$ have 1 in all columns of $\mathcal{F}$ [5, 6]. However, the major problem with frequent itemset is that if an itemset $\{a, b, c\}$ is frequent then their subsets are also frequent because of the anti-monotonicity property of frequent itemsets[7] thus making it unsuitable for comparison and reporting. On the other hand, maximal frequent itemset can be defined as an itemset which is frequent but non of its supersets are frequent [8]. Hence, we experiment our sampling methods with maximal frequent itemset.

UP'10, July 25th, 2010, Washington, DC, USA.

The rest of the paper is organized as follows. Section 2 briefly surveys the literature and Section 3 gives some brief information about the dataset used in these experiments. Section 4 and 5 discuss the methods used in upsampling and three methods used in downsampling. Section 6 summarizes the details of model selection procedure in the context of mixture models. Section 7 explains the implementation of the proposed methods, discusses the experiments performed on different methods, and also compares results of the experiments on different methods. In Section 8, conclusion is drawn from the experiments. Finally, Section 9 gives the future directions for research and issues not covered in this paper.

2. RELATED WORK

DNA copy number analysis was started in [9] where the authors mainly focused on determining the copy number of the cytogenetic band. Similar works performed were reviewed in [10] to determine the copy number. However, in [9] and [10], the authors did not establish a relation between the copy number and their clinical significance. In the recent past, DNA copy number amplification data collected with bibliomics survey from 838 journal articles published from 1992 to 2002 was analyzed in [11]. In the work, amplification patterns were determined for 73 different neoplasms and the neoplasms were clustered according to amplification profiles thus identifying the amplification hotspots using independent component analysis(ICA). The profiling revealed that human neoplasms formed clustered based on the amplification frequency. Continuing the studies in DNA copy number amplification, authors in [12] classified the human cancers based on copy number amplification using probabilistic modelling. Furthermore, the authors extracted the ranges of amplification in the chromosome and expressed it according to the cytogenetic nomenclature. In [13] and [14], the authors modeled the DNA copy number amplification using a mixture of multivariate Bernoulli Distribution. The classification of 73 different neoplasms in [11] were extended to 95 different neoplasm types. In [14] authors have proposed a compact and understadable reprsentation of the multivariate Bernoulli mixture model. Furthermore, in [15], the authors have proposed the enhancement to Bayesian Piecewise Constant Regression(BPCR) called mBPCR changing the segment number estimator and boundary estimator to enhance the fitting procedure. The proposed mBPCR was more accurate in the determination of true breakpoints of amplification. The more recent studies [16] and [17] have mainly focused in cancer specific analysis of DNA copy number. Although the mixture models were used in [13] and [14], they have studied only chromosome-1 data in resolution 393. Chromosome-1 being the largest chromosome, there are significant amount of amplifications [11]. However, a single chromosome band and the specific gene responsible for cancer has not been identified. Hence, study was performed on all chromosomes including chromosome-1. Chromosomewise analysis can reveal interesting facts about amplification of specific chromosomes and guarantees efficient computation & ease of analysis. Furthermore, there are several sources of multilevel biological data that comes in multiple resolutions but there seems to be a significant gap in research to deal with multiple resolution of the data. Algorithms and methods to deal with such multi-resolution data could possess very high clinical significance.

3. DATASET

The dataset provided was a binary (0-1) dataset about DNA amplifications specifying amplification of certain band of chromosome. DNA copy number amplifications are mutations in the DNA structure. The data was collected by bibliomics survey of 838 journal articles during 1992-2002 by hand without using state-of-the-art text mining techniques [12, 14]. The dataset contained the information about the amplification patterns of 4590 cancer patients. Each row describes one sample of cancer patient while each column identifies one chromosomal band(region). The amplified chromosomal regions were marked with 1 while and the value 0 defines that the chromosome band is not amplified. Chromosomes X and Y were not included in the experiments because of the lack of data. Patients whose chromosomal band had not shown any amplification for specific chromosome were not included in the experiments Thus different chromosomes had different number of samples.

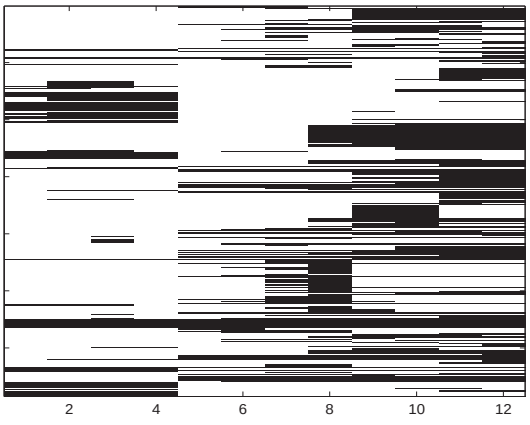

Figure 1: ***DNA copy number amplifications in chromosome-17, resolution 393.*** $\chi = (X_{ij})$, $X_{ij} \in \{0, 1\}$ **. Each row represents one sample of the amplification pattern for a patient and each column represents one of the chromosome bands.**

The data for chromosome-17 in resolution 393 demonstrated in the Figure 1, the copy number amplifications occur very sparsely and are often skewed. The original data was in the resolution 400 i.e. there were 393 chromosomal bands(regions) for the entire genome. The original data was upsampled to resolution 550, 700 and 850 and downsampled to resolution 300 using the methods discussed in Section 5. Bands for specific chromosome were extracted and mixture modelling was preformed on each chromosome. For example: chromosome-1 had 63, 61, 42, 28, and 23 chromosomal bands in resolution 850, 700, 550, 400, and 300 respectively [1]. Similarly, a different set of data was available in resolution 850. The data in resolution 850 was different than that in the resolution 400. Similar to the data in the resolution 400, the data in resolution 850 was downsampled to resolution 300, 400, 550 and 700. Element-wise AND operation over all the samples in the data results in a zero vector which necessitates sophisticated machine learning and data mining methods and techniques for classifying and profiling amplification.

4. UPSAMPLING

Upsampling is the process of changing the representation of data to the higher or finer resolution. A simple method was devised to upsample the data from resolution 393 and three different methods were used to downsample the data from higher resolution. Upsampling was simple and were implemented using simple transformation tables. Initially, the dataset was in resolution 393 and it was upsampled to three different resolutions 550, 700 and 850. A simple method was used to upsample the data. Multiple copies of cytogenetic band in lower resolution were made to upsample the data to higher resolution. For example, cytogenetic band 1q36.1 in resolution 550 has been divided into three bands 1q36.11, 1q36.12 and 1q36.13 in resolution 850. So, multiple copies of 1q36.1 was made for all bands 1q36.11, 1q36.12 and 1q36.13 in resolution. Figure 2 depicts the process of upsampling.

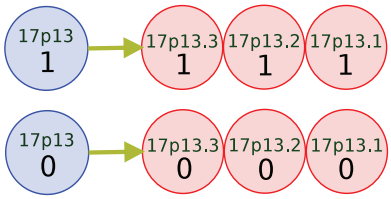

Figure 2: ***Schematic representation of upsampling where duplicate copies of similar cytogenetic bands are made in the higher resolution.***

Figure 2 shows that three copies of similar cytogenetic band in lower resolution band are made to upsample the data to higher resolution. When multiple copies of same cytogenetic band is made higher resolution will have only few unique rows. Hence, when the sample size decreases the complex model in higher dimension can not be trained to convergence thus producing poor results. Implementation of downsampling was performed using simple transformation tables implemented in Perl[18]. Table 1 shows an example of table for transformation of data in 393 resolution to 850 resolution for chromosome 17.

Table 1 shows that some chromosome bands missing in 393 resolution are seen in resolution 850. Hence, duplicate copies of the similar chromosome band in resolution 393 were made in higher resolution. Duplications were based on the assumption that if an adjacent area is amplified then the probability of the chromosome band being amplified is high because amplifications typically cover large areas. The transformation table were chromosome specific and resolution specific (i.e. 88 transformation table in all for different chromosomes)

5. DOWNSAMPLING

Downsampling is the process of changing the representation of the data to the lower or coarser resolution. In both cases of upsampling and downsampling no attempt is made to infer the structure of the data and no information is added or removed during the process. If the data of the same patients were available in two different resolutions, one of the supervised classification algorithm machine learning could be used in downsampling. However, such data was not available and hence simple but useful methods are used for downsampling. Downsampling methods were implemented in scripts with a script for each chromosome in each resolution. Section 5.1, 5.2 and 5.3 detail the methods of downsampling. Interestingly, in some cases there were some cytogenetic bands which were not available in higher resolution. For instance, the q arm of chromosome-4 in resolution 850 is divided into 4q35.1 and 4q35.2. In contrast, in resolution 700, the q arm of chromosome -4 is divided into three bands: 4q35.1, 4q35.2 and 4q35.3. respectively. In such cases, missing band in lower resolution was assigned the amplification pattern of its nearest neighbor in all three methods. For the example case above, the cytogenetic band 4q35.3 was assigned the amplification pattern of 4q35.2.

Resolution 393	Resolution 850
17p13	17p13.3
...	17p13.2
...	17p13.1
17p12	17p12
17p11.2	17p11.2
17p11.1	17p11.1
17q11.1	17q11.1
17q11.2	17q11.2
17q12	17q12
17q21	17q21.1
...	17q21.2
...	17q21.31
...	17q21.32
...	17q21.33
17q22	17q22
17q23	17q23.1
...	17q23.2
...	17q23.3
17q24	17q24.1
...	17q24.2
...	17q24.3
17q25	17q25.1
...	17q25.2
...	17q25.3

Table 1: Chromosome bands for resolution 393 & 850 and their transformation.

5.1 OR-function Downsampling

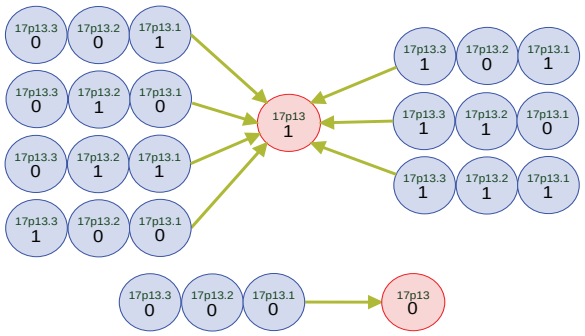

Figure 3: Schematic representation of OR-function downsampling procedure. Here the cytogenetic band in lower resolution is amplified if any of the bands in higher resolution is amplified. Cytogenetic band in lower resolution is not amplified only when none of the bands in higher resolution is amplified.

In OR-function downsampling method, the cytogenetic band in lower resolution is not amplified if none of the bands in higher resolution are amplified. The cytogenetic band in lower resolution is amplified if either of the bands in higher resolution is amplified. Figure 3 depicts the OR-function downsampling method. The OR-function downsampling method is based on simple belief that if the one of the bands in higher resolution is amplified, it signifies the presence of amplification in the band. For the case in the Figure 3 downsampling can be considered as a simple binary classification problem in machine learning where input is three dimensional binary variable and output is one dimensional binary variable. The solution is a simple truth table describing the classical OR operation.

5.2 Majority decision Downsampling

In majority decision downsampling method, a cytogenetic band in lower resolution is amplified if majority of the cytogenetic bands in higher resolution are amplified otherwise the cytogenetic band is not amplified. In case of a tie amplification of two nearest bands one in the left and other one in the right are taken into consideration iteratively and the amplification pattern of the band is determined using idea similar to 'golden goal'[1] strategy used in football. In other words, if in any iteration both bands in neighborhood bands are amplified than the band is amplified and if both the neighbors are unamplified than the band is deemed unamplified. If the amplification of lower resolution can not be concluded with 'golden goal' strategy then the band in lower resolution is deemed as amplified. Figure 4 shows one of the examples of majority decision in downsampling.

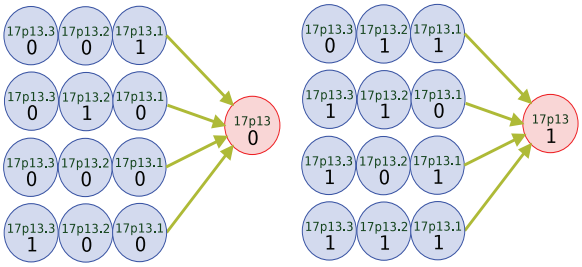

Figure 4: Schematic representation of majority decision downsampling procedure. Here the cytogenetic band in lower resolution is amplified if majority of the bands in higher resolution are amplified, otherwise it not amplified.

There is a shortcoming in this downsampling procedure because the majority decision downsampling procedure does not take into account the the lengths of the cytogenetic bands. The lengths of cytogentenic bands are considered by length weighted downsampling method discussed in Section 5.3.

5.3 Length weighted Downsampling

As shown in the Figure 5, length weighted downsampling method considers the length of the cytogenetic band. The

[1]The golden goal is a method used in football to determine the winner which end in a draw after the end of regulation time. Golden goal rules allow the team that scores the first goal during extra time to be declared the winner. The game finishes when a golden goal is scored.

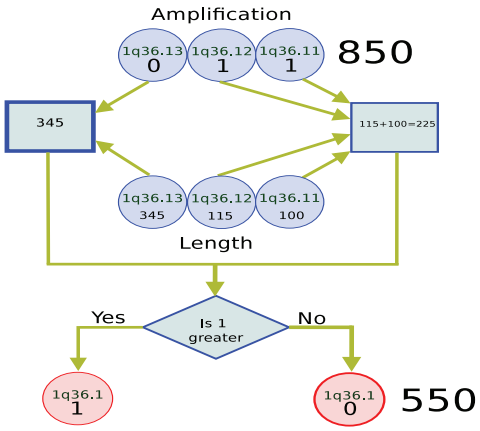

Figure 5: Schematic representation of weighted average downsampling procedure. Here the cytogenetic band in lower resolution is amplified if total length of the amplified bands in higher resolution is greater than the total length of unamplified bands, otherwise it not amplified. The figure is an example case in chromosome 1q36.1 where two cytogenetic bands 1q36.11 and 1q36.12 in resolution 850 are amplified and one band 1q36.13 is not amplified. However, total length of unamplified region i.e. band 1q36.13 (345) is greater than total length of the unamplified region i.e. bands 1q36.11 and 1q36.12 (100+115=225). Hence, the band in resolution 550 is unamplified.

length of the cytogenetic band varies in each assembly and hence relative lengths were considered. The amplification of cytogenetic band in lower resolution is determined by the weighted length of cytogenetic band in higher resolution. Each cytogenetic band is weighted according to the relative length of the cytogenetic band. If the total length of amplified region is greater than the total length of unamplified region, the cytogenetic band in lower resolution is amplified, otherwise the cytogenetic band is unamplified. Here, relative length is considered which gives more accurate measure of the amplification profiles in the cytogenetic band. Absolute lengths of the cytogenetic bands are not currently available and vary with each assembly. Two relative measures were considered in the calculation of the length. From the ideogram dataset available in NCBI [19], the difference between ISCN.top and ISCN.bot were used as relative measures. Similarly, difference between bases-top and bases-bot were also used as the relative measure of the length of each cytogenetic band. The difference in the results produced using the different relative measure of length have also been studied.

6. MODEL SELECTION

Cancer is not a single disease but a collection of diseases. Furthermore, cancer is a multi-factorial[2] disease. Therefore, finite mixture models [20, 21, 22] was selected to model the amplification data because they provide efficient method to

[2]Here multi-factorial is used to mean there are many factors causing cancer. Majority of the noninfectious diseases are multi-factorial.

model the heterogeneous population. Furthermore, since the copy number amplification data was a high dimensional binary data, the distribution used in the mixture model is Bernoulli distribution. Assuming that the data comes from a mixture of known number of components, J, finite mixture of multivariate Bernoulli distributions is defined as:

$$p(\mathcal{D}|\Theta) = \sum_{j=1}^{\mathrm{J}} \pi_j \prod_{i=1}^{\mathrm{d}} \theta_{ji}^{x_i} (1-\theta_{ji})^{1-x_i} \quad (1)$$

where π_j are the mixture proportions satisfying the properties such as convex combination such that $\pi_j \geq 0$ and $\sum_{j=1}^{\mathrm{J}} \pi_j = 1$ for all $j = 1, \ldots J$. Θ is composed of $\theta_1, \theta_2, \theta_3 \ldots \theta_d$ for each component distribution. Selection of number of mixture components J directly influences the performance of the mixture models. With fewer number of components, the mixture model behaves similar to a parametric model and increases the bias. On the contrary, if the mixture model has a large number of components then the model can overfit the data thus producing unreasonable variation. Hence, there is always a trade-off between the two. To optimize the trade-off and determine optimal number of components in the mixture model, 10-fold cross-validation technique [23, 24] was used. Expectation Maximization (EM) algorithm [25, 26] was used to train the mixture model using BernoulliMix programme package [27] freely available in BernoulliMix homepage. The model selection approach used in the paper is similar to [13, 14] except for the cross-validation procedure.

7. EXPERIMENTS

The downsampling methods were implemented in scripts, one each for each method, each chromosome and each resolution. Chromosomes X and Y were excluded from the experiments because of the lack of data. Hence, there were 198 scripts in all for all transformations. Matlab ®[28] was used for scripting. The individual scripts for downsampling each chromosome takes a file name of the data set in higher resolution as input checks for the abnormality in the data. The data was then transformed band-wise to lower resolution combining the multiple bands in higher resolution according to the three different methods proposed in Sections 5.1, 5.2 and 5.3. Furthermore, samples which contained no amplifications were also removed from the data.

7.1 Comparison of Downsampling Methods

The downsampled data from 850 resolution was subjected to various tests to determine the difference in the results of the downsampling methods. Few criteria were implemented to check the similarity of the results. Since we are working with data amplification patterns in cancer, the first difference measure used is the number of amplifications produced by the three downsampling methods. Total number of differences in each chromosome band was computed and compared between three different downsampling methods. Figure 6 depicts that the results of the three different downsampling process did not show significant differences with respect to the number of amplifications.

Scrutinizing the results further mean difference between the number of differences produced in the number of amplifications by the three methods in various chromosome bands was computed.

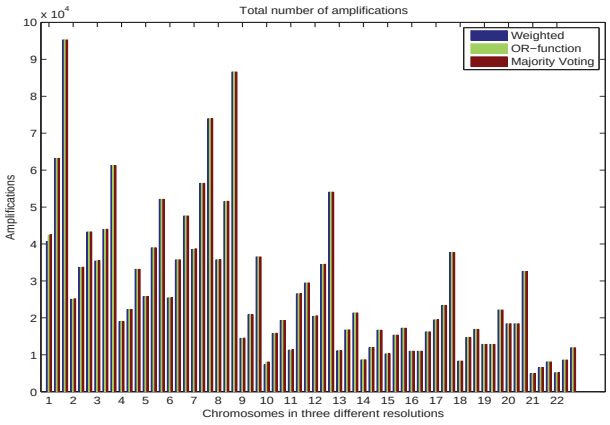

Figure 6: Total number of amplifications produced by the three different downsampling methods.

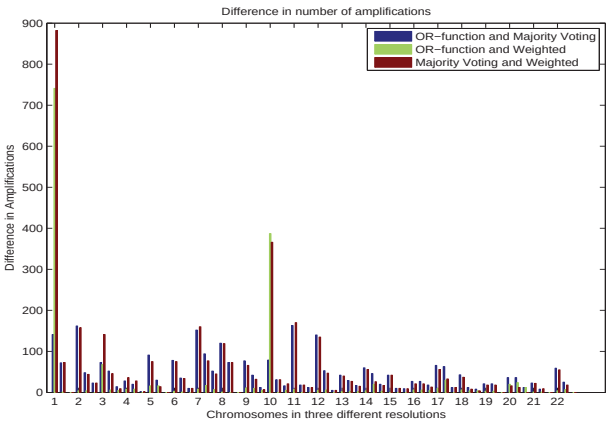

Figure 7: Difference in total number of amplifications produced by the three different downsampling methods.

Figure 7 suggests that there are differences in the results produced by the three downsampling methods with respect to the downsampling methods. However, the difference between the methods are not significant when the number of amplifications are considered. Similarly, other trivial difference measures such as row and column margins, and number of unique rows were also studied and the results showed that results of the downsampling methods are fairly similar.

However, these trivial measures used to calculate the difference are susceptible to some errors where the number of amplifications are same and also the number of amplifications does not change in different rows. For example, these methods does not show difference between the following two datasets.

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

In order to capture these differences, we further analyzed the difference between the two methods as the difference between the two resulting matrices for different methods using

standard matrix difference measures. The distance measure used is the square of the Frobenius norm [29] between two matrices. In binary matrices, Frobenius norm is essentially the number of cells where the two matrices differ.

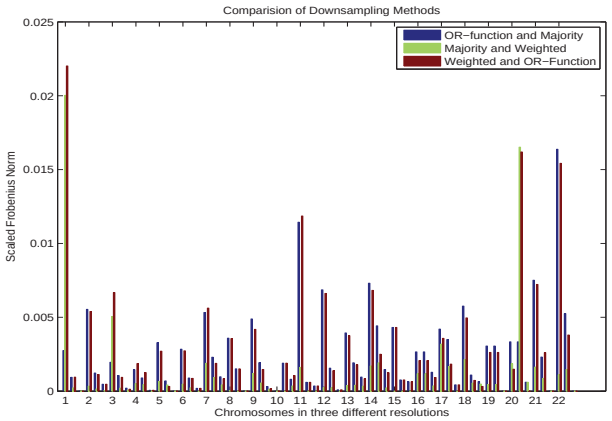

Figure 8: Comparison of three different downsampling methods : The difference measure used is scaled Frobenius norm.

Figure 8 suggests that the three downsampling methods produces fairly similar results. The Figure 8 also suggests that the differences are high in chromosome-1 which is expected because chromosome-1 is the largest chromosome. Differences are also high in lower resolution when compared to higher resolution because it is the lower resolution where the most changes takes place. The differences in the smaller chromosomes especially 20-22 are because of significant variation in the bands combined. Normally, three bands in finer resolution are combined in coarser resolution but in small chromosomes, the number of chromosome bands combined is very different thus making it difficult for weighted and OR-function downsampling method to work. It is to be noted that in the chromosomes where the differences are larger have larger number of differences in number of chromosome bands in different resolutions.

7.2 Model selection in Mixture Model

The size of the chromosome in terms of chromosome bands varied significantly. Some chromosomes had higher number of bands and some chromosomes had lower number of chromosome bands. Data from different resolutions were individually subjected to the mixture models. For model selection, for each mixture component, 50 models were trained using training set. It is often recommended to repeat cross-validation technique a number of times because 10-fold cross-validation can be seen as a "standard" measure of the performance whereas ten 10-fold cross-validations would be a "precise" measure of performance [30]. Since EM-algorithm is sensitive to the initializations and the results may differ on the same data for different initializations and it can get stuck in local minima and the global optimum results are not often guaranteed [31], 50 different models were trained for each number of components. In other words, 10-fold cross-validation was repeated 50 times. The number of mixture components was varied from 2 to 20 for all chromosomes in all resolutions. Validation set for each model is the one remaining subset of the data which is not used for training. Total likelihood for the training data as well as the validation data is calculated and averaged for each mixture component. The number of components for which the likelihood is maximum is selected as the model for the data taking parsimony into account. In other words, in some cases, models with lesser mixture components are selected instead of models with large number of mixture components for which likelihood was higher. Model selection was performed on all chromosomes as chromosome-wise analysis can reveal interesting facts about amplification of specific chromosomes and guarantees efficient computation & ease of analysis. Here, results are explained only for chromosome-17 as an example. Two sets of original data were available in resolution 393 and 850. Experiments were performed in the original resolution and sampling was performed to sample the data to different resolutions. Experiments showed that number of components required to optimally fit the model is independent of the resolution of the data thus showing that the significant patterns are not lost during sampling. Figure 9 and 10 show a model selection procedure for the data in resolution 393 and 850.

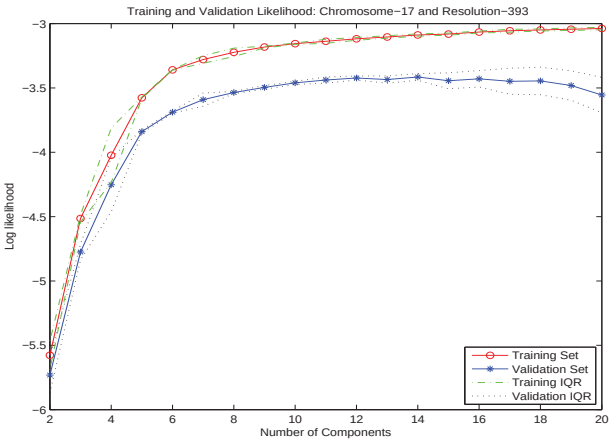

Figure 9: Model selection for the the original data in resolution 393. The averaged log-likelihood for training and validation sets in a 10-fold cross-validation setting for different number of components in chromosome-17: Resolution-393. The interquartile range(IQR) for 50 different training and validation runs have also been plotted.

Figure 9 shows the model selection in case of resolution 393 which downsampled from resolution 850. Figure 9 shows that the likelihood is smoothly increasing function with respect to the number of components. From Figure 9, it can be seen that validation likelihood is maximum when the number of components is 14, but instead of 14 components, 6 components was selected. It is to be noted that sometimes complex models overfit the data. Simple model also reduces the time and space complexity. Furthermore, the training and validation likelihood when the number of components is 6 are -3.3593 and -3.6883. In addition, when the number of components is 14, the training and validation likelihood are -3.0887 and -3.4146. Hence the difference in likelihood is negligible when compared with the efficiency in terms of time and space complexity. Furthermore, when the number of components are increased, IQR shows significant varia-

tion. The variation in IQR is because when the number of components are increased, samples can be assigned to different clusters. Additionally, the data in resolution in 393 was upsampled to resolution 850 and similar approach for model selection was followed.

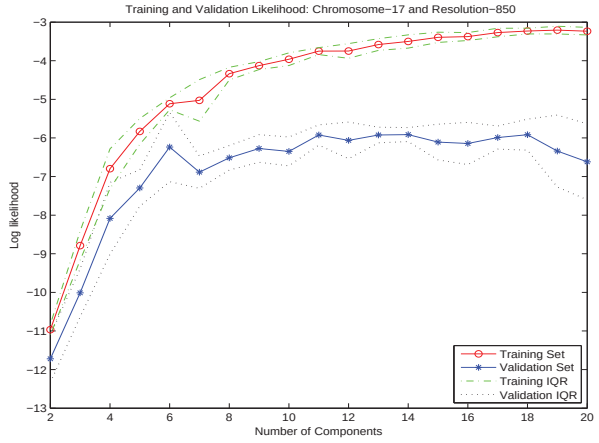

Figure 10: The averaged log-likelihood for training and validation sets in a 10-fold Cross validation setting for different number of components in chromosome17: Resolution 850. The interquartile range(IQR) for 10 different training and validation runs have also been plotted.

Figure 10 also shows that the IQR varies significantly from the mean. The choice of the number of components is straightforward because Figure 10 clearly shows a maximum of validation likelihood when the number of components is 8. Even when the number of components is 8, the variation in IQR is high. The variation in IQR can be compensated with sufficient training and would produce favorable results. The results can be further improved when the size of the dataset is increased.

The major aim of upsampling and downsampling was to aid in the integration of databases. The clinical aspects regarding the classification of cancer with mixture models is already established in [11] and [12]. Thus, data in different resolution were combined after upsampling and downsampling and model selection was performed. Table 2 summarizes the results of the experiments on chromosome: 17. To calculate the Likelihood 50 different models were trained to convergence and likelihood of the data was calculated for each model and the mean of the results are reported.

Data Resolution	J	Likelihood
Original in 393	8	-3.39
Original in 850	8	-4.75
Downsampled to 393	6	-3.41
Upsampled to 850	6	-5.23
Combined in 393	7	-3.36
Combined in 850	7	-5.11

Table 2: Results of experiments on chromosome-17. J denotes the selected number of component distributions.

Table 2 shows the number of components required to fit the data differs in different resolution and different number of samples in the data. The likelihood of data in higher resolution is lower than the likelihood of the data in the lower resolution when the number of components are same. This phenomenon can be attributed to the curse of dimensionality [32]. For example, the dimensionality of data in resolution 393 and 850 differs by 12 in chromosome-17 but likelihood is lesser even when the number of components is similar. For the original data in resolution 393 and 850, the difference in number of parameters of the model is $6 * (1 + 26) - 6 * (1 + 18) = 48$ which invites significant amount of computational complexity. The increased complexity however does not produce corresponding the increase in the likelihood. With increasing samples, the number of components are not increased because the complexity of mixture models depends on the complexity of the problem being solved, not with the size of dataset. This experiments with the mixture models also shows that patterns present in the higher resolution of the data is efficiently and effectively preserved in lower resolution.

Data Resolution	# X	Train	Test
Original in 393	342	0.25	0.06
Original in 850	2716	0.43	0.30
Downsampled to 393	2716	1.12	0.20
Upsampled to 850	342	2.16	0.08
Combined in 393	3058	1.43	0.19
Combined in 850	3058	2.51	0.32

Table 3: Computational complexity for training and testing of a single mixture model with appropriate number of mixture components as decided in 2. Experiments are performed on chromosome-17 and time is calculated in seconds. X denotes the number of data samples. The hardware used is Intel Core2Duo 2.00GHz CPU with a memory of 3 GB.

The major drawback in using mixture models is computational complexity of training the mixture models. Normally, training mixture models are computationally expensive when compared to other parametric (such as Poisson distribution) as well as non-parametric (such as k-means) methods. Similar to other machine learning methods computational complexity of the mixture model also increases with increasing dimension i.e. resolution in our case. Thus, computational complexity was also estimated for each resolution for the number of components shown in Table 2. As shown in the Table 3 the computational complexity increases with increasing resolution. To estimate the training time, 50 different models are trained until 10 iterations and the mean of the result is taken as final training time. Similarly, likelihood is calculated for 50 different models trained to calculate the training time and the mean of the results is reported. Experiments with resolution 850 required approximately twice the time required for the resolution 393. Furthermore, from Table 2, we also know that number of components required is high when the resolution is increased but the likelihood decreases. In addition, the curves are smoother in Figure 9 when compared to Figure 10. This phenomenon is because of the intrinsic problems of working with high dimensional data arising in higher resolution. These results suggest that data in lower resolution is preferred but lower resolution does not capture all the available biological information. Thus, there is a trade-off between the two.

7.3 Frequent itemsets

The measure of frequent itemsets provides a metric for the similarity measure between the sampled data and original data. Furthermore, our major aim was to upsample and downsample the data so that the patterns in the original resolution were retained. Mining maximal frequent itemset in the context of mixture modelling of multivariate Bernoulli distribution is two fold. It has by shown in [14] that maximal frequent itemset can be used to describe the finite mixture of multivariate Bernoulli distributions compactly and in a language understandable by the domain experts. In [14], the authors implemented a mixture of Bernoulli distributions in clustering binary data to derive frequent itemsets from the cluster-specific data sets and found that the cluster-specific maximal frequent itemset were significantly different from those itemsets extracted globally.

Similar to [14], we used MAFIA (MAximal Frequent Itemset Algorithm) [8] to mine the frequent patterns because other similar algorithms such as Apriori [6] would produce long results which will be difficult to interpret. The frequency or the threshold was choosen as 0.5 motivated by a majority voting protocol. Upscaling is simple and is always guaranteed to retain the frequent itemset although the number of frequent itemset increases with the exact same support. Therefore, they have not been reported.

Data	Maximal frequent itemsets
Og. 393	{11},{12}
Og. 850	{7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24}
OR. 393	{5, 6, 7, 8, 9, 10, 11, 12}
We. 393	{7,8}, {5, 6, 7}, {7,12}, {7,11}, {8, 9, 10, 11, 12}
Mj. 393	{5, 6, 7, 8, 9, 10, 11, 12 }
Co. 393	{5, 6, 7}, {6,7,8}, {7, 8, 9, 10, 11}, {7, 8, 11, 12}, {8, 9, 10, 11, 12}
Co. 850	{7, 8, 9}, {8, 9, 10, 11, 12, 13, 14}, {9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21}, {9, 10, 11, 12, 13, 14, 19, 20, 21, 22, 23, 24 }, { 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24}

Table 4: Maximal frequent itemsets for data in different resolutions. The threshold used is 0.5. Og., OR., We., Mj. and Co. denotes the original, OR-function down-sampled, weighted down-sampled, majority voting downsampled and combined data respectively.

From Table 4, we can see that the maximal frequent itemsets are preserved during sampling of resolutions. For example, in OR-function downsampled data in resolution 393 and original data in resolution 850, there is no difference in the maximal frequent itemset because from upsampling Table 1 we know that items 7,8, and 9 in 850 represents items 5, 6 and 7. Items 8 to 14 in 850 are combined to form item 8 in the data. Other itemsets are also formed with similar combinations. Weighted downsampling differs more than other two types of methods but the difference is not significant. The results of sampling can be seen more profoundly in integrated datasets where each itemsets in higher resoulution can be defined by the frequent itemsets lower resolution. The differences in some cases are only seen because support for those itemsets are less; these differences can be expected because data in lower resolution can not encompass all the information in higher resolution.

8. SUMMARY AND CONCLUSIONS

A simple upsampling and three different downsampling methods were proposed and their results were studied. The results were plausible and fairly consistent. The resulting data in different resolutions efficiently captures the information of data in different resolutions. Mixture models were then applied to the data in different resolutions. Finally, data in two different resolutions were integrated and then analyzed in one resolution. The results suggested that number of components required to fit the data does not differ across resolutions but likelihood of the model on higher resolution is poor than on lower resolution although the data is the same but representation is different. The clustering results of mixture models possesses high clinical significance. Furthermore, the maximal frequent itemsets and mixture modelling show that significant patterns in the data is maintained during sampling.

9. FUTURE WORK

Mixture models are limited because they work with one resolution of data. In the future work, they can be extended to work with multiple resolutions of the data where the sampling is incorporated with in models. The sampling techniques can be constrained to maintain the significant patterns in the dataset.

10. REFERENCES

[1] L.G. Shaffer and N. Tommerup. *ISCN 2005: An International System for Human Cytogenetic Nomenclature(2005) Recommendations of the International Standing Committee on Human Cytogenetic Nomenclature*. Karger, 2005.

[2] A. Kallioniemi, O.P. Kallioniemi, D. Sudar, D. Rutovitz, J.W. Gray, F. Waldman, and D. Pinkel. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *SCIENCE*, 258(5083):818–821, OCT 30 1992.

[3] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W.L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B.M. Ljung, J.W. Gray, and D.G. Albertson. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20: 207 – 211, 1998.

[4] I. K. Fodor. A survey of dimension reduction techniques. Technical report, U.S. Department of Energy, June 2002.

[5] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, New York, NY, USA, 1993. ACM.

[6] H. Mannila, H. Toivonen, and A. I. Verkamo. Efficient algorithms for discovering association rules. In Usama M. Fayyad and Ramasamy Uthurusamy, editors, *AAAI Workshop on Knowledge Discovery in*

Databases (KDD-94), pages 181–192, Seattle, Washington, 1994. AAAI Press.

[7] Arianna Gallo, Pauli Miettinen, and Heikki Mannila. Finding subgroups having several descriptions: Algorithms for redescription mining. In *SDM*, pages 334–345, 2008.

[8] Doug Burdick, Manuel Calimlim, and Johannes Gehrke. Mafia: A maximal frequent itemset algorithm for transactional databases. In *In ICDE*, pages 443–452, 2001.

[9] J.R. Pollack, C.M. Perou, A.A. Alizadeh, M.B. Eisen, A. Pergamenschikov, C.F. Williams, S.S. Jeffrey, D. Botstein, and P.O. Brown. Genome-wide analysis of dna copy-number changes using cdna microarrays. *Nature Genetics*, 23(1):41–46, 1999.

[10] S. Knuutila, Y. Aalto, K. Autio, A. Björkqvist, W. El-Rifai, Hemmer S., T. Huhta, E. Kettunen, S. Kiuru-Kuhlefelt, M.L. Larramendy, T Lushnikova, O. Monni, H. Pere, J. Tapper, M. Tarkkanen, A. Varis, V. Wasenius, M. Wolf, and Y. Zhu. Dna copy number losses in human neoplasms. *Gynecologic Oncology*, 155(2):683–694, 1999.

[11] S. Myllykangas, J. Himberg, T. Böhling, B. Nagy, J. Hollmén, and S. Knuutila. DNA copy number amplification profiling of human neoplasms. *Oncogene*, 25(55):7324–7332, 2006.

[12] S. Myllykangas, J. Tikka, T. Böhling, S. Knuutila, and J. Hollmén. Classification of human cancers based on DNA copy number amplification modeling. *BMC Medical Genomics*, 1:15, 2008.

[13] J. Tikka, J. Hollmén, and S. Myllykangas. Mixture modeling of DNA copy number amplification patterns in cancer. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4507 LNCS:972–979, 2007.

[14] J. Hollmén and J. Tikka. Compact and understandable descriptions of mixtures of bernoulli distributions. *Lecture Notes in Computer Science including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, 4723 LNCS:1–12, 2007.

[15] P.M.V. Rancoita, M. Hutter, F. Bertoni, and I. Kwee. Bayesian DNA copy number analysis. *BMC Bioinformatics*, 10, 2009.

[16] B. D'haene, J. Vandesompele, and J. Hellemans. Accurate and objective copy number profiling using real-time quantitative PCR. *Methods*, 50(4):262–270, 2010.

[17] E. Despierre, D. Lambrechts, P. Neven, F. Amant, S. Lambrechts, and I. Vergote. The molecular genetic basis of ovarian cancer and its roadmap towards a better treatment. *Gynecologic Oncology*, 117(2):358–365, 2010.

[18] L. Wall. Perl: Practical Extraction and Report Language. Website, 1987. http://www.perl.org/: Last Accessed: 15 Mar 2010.

[19] National Center for Biotechnology Information. Human genome project. Website, February 2010. `http://www.ncbi.nlm.nih.gov/projects/mapview/` Last Accessed: 5 Feb 2010.

[20] G. J. McLachlan and D. Peel. *Finite mixture models*, volume 299 of *Probability and Statistics – Applied Probability and Statistics Section*. Wiley, New York, 2000.

[21] B. S. Everitt and D. J. Hand. *Finite mixture distributions*. Chapman and Hall, 1981.

[22] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1st ed. 2006. corr. 2nd printing edition, October 2007.

[23] S. Geisser. A predictive approach to the random effect model. *Biometrika*, 61(1):101–107, 1974.

[24] F. Monsteller and J. Tukey. Data analysis including statistics. *In Lindzey G. and Aronson E., editors, Handbook of Social Psychology, Vol-2, Addison-Wesley*, 1968.

[25] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal Of The Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[26] J. H. Wolfe. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5:329–350, 1970.

[27] J. Hollmén. *BernoulliMix: Program package for finite mixture models of multivariate Bernoulli distributions*, May 2009. Freely available in http://www.cis.hut.fi/jHollmen/BernoulliMix/.

[28] Mathworks. Matlab: the language of technical computing. Website, 1994. http://www.mathworks.com/products/matlab/: Last Accessed: 15 Mar 2010.

[29] G. W. Stewart. *Matrix Algorithms: Volume 1, Basic Decompositions*. Society for Industrial Mathematics, 1998.

[30] S.D. Gay. *Datamining in proteomics: extracting knowledge from peptide mass fingerprinting spectra*. PhD thesis, University of Geneva, Geneva, 2002.

[31] G. J. Mclachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley-Interscience, 1 edition, November 1996.

[32] W. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley, 2007.

Publication II

Prem Raj Adhikari, Jaakko Hollmén. Fast Progressive Training of Mixture Models for Model Selection. ***Journal of Intelligent Information Systems*****, IN PRESS, Springer, DOI: 10.1007/s10844-013-0282-3, Published Online: December 2013.**

J Intell Inf Syst
DOI 10.1007/s10844-013-0282-3

Fast progressive training of mixture models for model selection

Prem Raj Adhikari · Jaakko Hollmén

Received: 11 January 2013 / Revised: 15 September 2013 / Accepted: 4 October 2013

Abstract Finite mixture models (FMM) are flexible models with varying uses such as density estimation, clustering, classification, modeling heterogeneity, model averaging, and handling missing data. Expectation maximization (EM) algorithm can learn the maximum likelihood estimates for the model parameters. One of the prerequisites for using the EM algorithm is the a priori knowledge of the number of mixture components in the mixture model. However, the number of mixing components is often unknown. Therefore, determining the number of mixture components has been a central problem in mixture modelling. Thus, mixture modelling is often a two-stage process of determining the number of mixture components and then estimating the parameters of the mixture model. This paper proposes a fast training of a series of mixture models using progressive merging of mixture components to facilitate model selection algorithm to make appropriate choice of the model. The paper also proposes a data driven, fast approximation of the Kullback–Leibler (KL) divergence as a criterion to measure the similarity of the mixture components. We use the proposed methodology in mixture modelling of a synthetic dataset, a publicly available zoo dataset, and two chromosomal aberration datasets showing that model selection is efficient and effective.

Keywords Model selection · Mixture models · KL divergence · Training · 0–1 data

P. R. Adhikari ( ) · J. Hollmén
Helsinki Institute for Information Technology (HIIT),
Department of Information and Computer Science (ICS),
Aalto University School of Science,
PO Box 15400, 00076 Aalto, Espoo, Finland
e-mail: prem.adhikari@aalto.fi

J. Hollmén
e-mail: jaakko.hollmen@aalto.fi

Published online: 01 December 2013

1 Introduction

Finite mixture models are flexible probabilistic models suitable for modelling complex data distributions. They have varying uses such as density estimation, clustering, classification, model averaging, and handling missing data (McLachlan and Peel 2000; Everitt and Hand 1981). EM algorithm provides a conceptual framework to estimate the maximum likelihood parameters from incomplete data (Dempster et al. 1977). Formulation of the EM algorithm provided the necessary impetus to the growing use of mixture models. Estimation of the parameters of mixture models involves numerous challenges. One of those challenges is the requirement of a priori knowledge of the number of components in the mixture model. Model selection in mixture model is the method of selecting the appropriate number of components in a mixture model (Tikka et al. 2007). It is one of the central problems in the finite mixture modelling.

A large number of mixture components fit the data better producing a high likelihood values for the training set. However, it also increases the model complexity, and results in an over-fitted model. The over-fitted model often generalizes poorly on future data. Conversely, smaller number of mixture components may result in an under-fitted model, which provides poor accuracy in modeling. Often some validation methods optimize this trade-off between the model accuracy and the generalization ability. However, mixture models are unsupervised models making it difficult to determine the error measure which is most widely used during validation.

Data likelihood is widely used to compare and determine the generative performance of mixture models in an unsupervised setting using cross-validation (Smyth 2000). Authors have proposed different deterministic, and stochastic and re-sampling methods to estimate the number of components in a mixture model (Figueiredo and Jain 2002). Deterministic methods consist of methods such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Minimum Description Length (MDL). Stochastic and re-sampling methods consist of methods such as Markov Chain Monte Carlo (MCMC). We have used cross-validated likelihood to choose the number of components in a mixture model to model the copy number aberration patterns in cancer patients in our previous work in Tikka et al. (2007), Hollmén and Tikka (2007) and Adhikari and Hollmén (2010a, b). Similarly, in Ueda et al. (2000), the authors used splitting and merging of mixture components to ameliorate the problem of local optima in the EM algorithm. Furthermore, Zhang et al. (2003) presented another split and merge algorithm which uses different split and merge criterion such as Singular Value Decomposition (SVD) and Cholesky decomposition to split and merge components. In Ueda et al. (2000) and Zhang et al. (2003), the authors used a fixed number of components and repeated splitting and merging on the same number of mixture components to search for the global optimum.

Authors use split and merge strategy in conjunction with a validation method to select number of components in a mixture model. For example, in Li and Li (2009a), the authors used MDL criterion with split and merge algorithm to select the number of components in mixture model. Similarly, Zhang et al. (2004) proposed a competitive EM algorithm that automatically learns the number of mixture components and is insensitive to the initial configuration of number of mixture components and model parameters. Similarly, Blekas and Lagaris (2007) proposed an optimization strategy

to determine the optimal number of mixing components using repeated split and merge operations. Model selection based on AIC (Akaike 1974) also uses the KL divergence as a criterion to select the number of components (Windham and Cutler 1992). In Windham and Cutler (1992), authors used the KL divergence between two prospective models penalizing the model with the higher number of mixture components. Some other distance measures proposed on measuring the dissimilarity between two Hidden Markov Models (HMMs) such as the one used in Juang and Rabiner (1985) are also based on the symmetric version of the KL divergence.

In all of these and the family of related methods, usually considered merge criteria is the similarity between the posterior probabilities of the component distributions. Ueda et al. (2000) used the normalized Euclidean distance between the two component distributions as a merge criterion. However, since the component distributions are probability distributions, using a geometric distance measure such as the Euclidean distance is unsuitable. Similarly, Zhang et al. (2004), Blekas and Lagaris (2007) and Li and Li (2009b) use the local KL divergence to measure the distance between the local data density and the model density of the component. However, the two component distributions are probability distributions and the local KL divergence provides unsatisfactory measure of the difference between them. The use of the full KL divergence is also restricted by its computationally expensive calculation.

Generally, the KL divergence between the two densities does not have a closed-form solution. Hence, authors have proposed different approximations of the KL divergence in the literature (Goldberger et al. 2003; Hershey and Olsen 2007). Similar to our method, approximation in Goldberger et al. (2003) and Hershey and Olsen (2007) are also based on the data re-sampling approach. However, they base their re-sampling approach on the MCMC while our assumption is that the samples in the data are the true samples of the distribution. Furthermore, most of the methods consider the Gaussian mixture model, and have adapted the algorithm to suit this particular choice of distribution. In our experiments, we use the finite mixture models of the multivariate Bernoulli distributions to model the chromosomal aberration patterns in cancer.

Authors have proposed different data driven approximations of the KL divergence. Lee and Park (2006) proposed two estimators of the KL divergence by local likelihood. Leonenko et al. (2008) proposed a technique to estimate the KL divergence using the Monte-Carlo estimator. Similarly, Wang et al. (2005) proposed a data–driven universal estimator of the divergence but only for continuous functions. Additionally, Perez-Cruz (2008) also proposed method for estimating the KL divergence between two continuous densities without the need for estimating the densities. These above methods are suitable for continuous densities. Authors have also proposed methods for discrete data. For instance, Cai et al. (2006) proposed two algorithms borrowed from data compression techniques to estimate the divergence from the realizations of two unknown finite–alphabet sources. Furthermore, our proposed approximation satisfies all three properties of the KL divergence but it is unusable as a metric because similar to the full KL divergence it dissatisfies triangle inequality.

This paper is the extension of our earlier paper (Adhikari and Hollmén 2012). In this paper, we propose an approximation of the exact KL divergence to determine the similarity between the mixture components with an aim to merge the most similar

components. We do not propose a model selection criterion and our series of models works in conjunction with any model selection criterion such as AIC, BIC, and MDL. We repeatedly merge mixture components and retrain the mixture models to generate a series of mixture models. For example, using cross-validation on the series of models, we can select the model with optimal number of components expected to produce the best generalization performance. We also propose a fast data driven approximation of the KL divergence and use it to select two candidate components to merge in a mixture model. We perform the experiments on different 0–1 datasets showing that the results of our methods are plausible.

The organization of current paper is as follows. Section 2 briefly reviews the mixture models of the multivariate Bernoulli distributions and the EM algorithm. Section 3 presents the KL divergence and its derivation for the finite mixture model of multivariate Bernoulli distributions to compare two mixture components in a mixture model. Section 4 discusses the experiments performed on the one synthetic dataset, one publicly available zoo dataset, and two real–world datasets describing chromosomal aberrations patterns in cancer and analyzes the obtained results. Section 5 draws the conclusions from the experimental results.

2 Mixture models and EM algorithm

Finite mixture models represent a statistical distribution using a mixture (or weighted sum) of simple distributions such as Gaussian, Poisson, and Bernoulli. We achieve this by decomposing the probability density function into a set of component density functions (McLachlan and Peel 2000; Everitt and Hand 1981). $\Theta = \{J, \{\pi_j, \theta_j\}_{j=1}^{J}\}$ parameterizes a finite mixture of multivariate Bernoulli distributions having J components for a data set $\boldsymbol{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ consisting of data vectors of dimensionality d. Mathematically, we define mixture models as:

$$p(\mathbf{x} \mid \Theta) = \sum_{j=1}^{\mathrm{J}} \pi_j \prod_{i=1}^{\mathrm{d}} \theta_{ji}^{x_i} (1 - \theta_{ji})^{1-x_i}. \tag{1}$$

Here π_j are the mixture proportions satisfying the properties such as convex combination: $\pi_j \geq 0$ and $\sum_{j=1}^{\mathrm{J}} \pi_j = 1 \ \forall \ j = 1, \ldots, J$. θ_{ji} defines the probability that random variable of the jth component in the ith dimension will take the value 1. Similarly, Θ_j denotes the vector of random variables of the component j. Therefore, $\Theta_j = \theta_{j,1}, \theta_{j,2}, \theta_{j,3}, \ldots, \theta_{j,d}$ where d denotes the dimensionality of the data. Θ denotes all the parameters of the data including mixture components. Therefore, $\Theta = \{J, \{\pi_j, \theta_j\}_{j=1}^{J}\}$. x_i denotes the data point such that $x_i \in \{0, 1\}$. Learning the parameters of a mixture model of Bernoulli distributions means learning the parameters Θ which includes the number of components J from the given data $\boldsymbol{X}$ of dimensionality d. We can formulate the learning in log-likelihood terms as:

$$\mathcal{L}(\Theta) = \sum_{n=1}^{\mathrm{N}} \log \ P(x_n \mid \Theta) = \sum_{n=1}^{\mathrm{N}} \log \left[\sum_{j=1}^{\mathrm{J}} \pi_j \prod_{i=1}^{\mathrm{d}} \theta_{ji}^{x_{ni}} (1 - \theta_{ji})^{1-x_{ni}} \right]. \tag{2}$$

Component-wise differentiation of (2) with respect to results θ_j and π_j in:

$$\frac{\delta \mathcal{L}}{\delta \pi_j} = \frac{1}{\pi_j} \sum_{n=1}^{\mathrm{N}} P(j|x_n; \pi, \Theta) - N \quad j = 1, \ldots, J \tag{3}$$

And also

$$\frac{\delta \mathcal{L}}{\delta \theta_{ji}} = \frac{1}{\theta_{ji}(1-\theta_{ji})} \sum_{n=1}^{\mathrm{N}} P(j|x_n; \pi, \Theta)(x_{ni} - \theta_{ji})$$
$$\text{where} \quad j = 1, \ldots, J \text{ and } i = 1, \ldots, d \tag{4}$$

The term -N in equation satisfies the constraint $\sum_{j=1}^{J} \pi_j$ introduced in loglikelihood via Lagrange multiplier. Now, from Bayes' theorem, we can calculate the posterior probability as:

$$P(j \mid \boldsymbol{X}; \Theta) = \frac{\pi_j \prod_{i=1}^{d} \theta_{ji}^{x_{ni}} (1-\theta_{ji})^{1-x_{ni}}}{\sum_{j'=1}^{J} \prod_{i=1}^{d} \theta_{j'i}^{x_{ni}} (1-\theta_{j'i})^{1-x_{ni}}}. \tag{5}$$

Now, the EM algorithm (Dempster et al. 1977; Wolfe 1970) is two stage iterative algorithm defined by:

E-step: E-step computes the posterior probability using (5) for the most recent values of parameters $\pi^{\tau}, \Theta^{\tau}$ at iteration τ, i.e., calculates $P(j \mid x_n; \pi^{\tau}, \Theta^{\tau})$

M-step: M-step recomputes the values of parameters $\pi^{\tau+1}, \Theta^{\tau+1}$ for the next iteration.

$$\pi_j^{(\tau+1)} = \frac{1}{N} \sum_{n=1}^{N} P\left(j \mid x_n; \pi^{(\tau)}, \Theta^{(\tau)}\right)$$

$$\Theta_j^{(\tau+1)} = \frac{1}{N\pi_j^{(\tau+1)}} \sum_{n=1}^{N} P\left(j \mid x_n; \pi^{(\tau)}, \Theta^{(\tau)}\right) x_n \tag{6}$$

Iterations between the E and the M steps produce a succession of monotonically increasing sequence of log-likelihood values for the parameters $\tau = 0, 1, 2, 3, \ldots$ regardless of the starting point $\{\pi^{(0)}, \Theta^{(0)}\}$ (Mclachlan and Krishnan 1996). The EM algorithm is sensitive to initializations but is deterministic for a given initialization and a given dataset.

3 Kullback–Leiber divergence

The Kullback–Leibler divergence is the non-symmetric difference between two probability distributions (Kullback and Leibler 1951; Cover and Thomas 1991). Mathematically, the KL divergence between two discrete probability distributions P, and Q on a finite set $\boldsymbol{X}$ is:

$$\mathcal{D}_{KL}(P \,||\, Q) = \sum_{x \in \boldsymbol{X}} P(x) \log \frac{P(x)}{Q(x)} \tag{7}$$

KL divergence is unsymmetric because KL divergence from P to Q is different from the KL divergence from Q to P. Nevertheless, we can symmetrize the KL divergence by summing the KL divergence from P to Q and from Q to P (Juang and Rabiner 1985). Symmetrized KL divergence satisfies the properties of distance metric such as positivity, self–similarity, and self-identification. However, KL divergence does not satisfy triangle inequality. Both the KL divergence and the EM algorithm work on the same quantity of likelihood. However, we are using the KL divergence to compare the two component distributions and not two prospective mixture models. We write the symmetrized KL divergence and reorder the terms to separate the difference of probabilities and the log likelihood ratio of the models as:

$$\begin{aligned}\mathcal{D}_{KL} &= \mathcal{D}_{KL}(P \,||\, Q) + \mathcal{D}_{KL}(Q \,||\, P) \\ &= \sum_{x \in \boldsymbol{X}} P(x) \log \frac{P(x)}{Q(x)} + \sum_{x \in \boldsymbol{X}} Q(x) \log \frac{Q(x)}{P(x)} \\ &= \sum_{x \in \boldsymbol{X}} \left\{ P(x) \log \frac{P(x)}{Q(x)} + Q(x) \log \frac{Q(x)}{P(x)} \right\} \\ &= \sum_{x \in \boldsymbol{X}} \left\{ (P(x) - Q(x)) \log \frac{P(x)}{Q(x)} \right\}. \end{aligned} \tag{8}$$

3.1 KL divergence between components of FMM of multivariate Bernoulli distributions

Let us denote the two component distributions, first P and second Q, from a mixture model as θ and β respectively. Also, let θ_k and β_k denote the k^{th} parameters of the component distributions θ and β respectively. Similarly, $\boldsymbol{X}$ is a matrix of random variables (x_{ik}) that denotes the binary state-space for the random variable x of given dimensionality d indexed by k. From Adhikari and Hollmén (2012), we can write the symmetric KL divergence, generalized to an arbitrary dimension of data d, for two component distributions in a mixture model as:

$$\begin{aligned} KL_{\theta\beta} = \sum_{i=1}^{2^d} \Bigg[\left\{ \prod_{k=1}^{d} \left(\theta_k^{x_{ik}} (1-\theta_k)^{(1-x_{ik})} \right) - \prod_{k=1}^{d} \left(\beta_k^{x_{ik}} (1-\beta_k)^{(1-x_{ik})} \right) \right\} \\ \cdot \log \prod_{k=1}^{d} \frac{\theta_k^{x_{ik}} (1-\theta_k)^{(1-x_{ik})}}{\beta_k^{x_{ik}} (1-\beta_k)^{(1-x_{ik})}} \Bigg]. \end{aligned} \tag{9}$$

We can replace the log and the product in the last term with a summation and log resulting in an equation of the form:

$$\begin{aligned} KL_{\theta\beta} = \sum_{i=1}^{2^d} \Bigg[\left\{ \prod_{k=1}^{d} \left(\theta_k^{x_{ik}} (1-\theta_k)^{(1-x_{ik})} \right) - \prod_{k=1}^{d} \left(\beta_k^{x_{ik}} (1-\beta_k)^{(1-x_{ik})} \right) \right\} \\ \cdot \sum_{k=1}^{d} \log \frac{\theta_k^{x_{ik}} (1-\theta_k)^{(1-x_{ik})}}{\beta_k^{x_{ik}} (1-\beta_k)^{(1-x_{ik})}} \Bigg]. \end{aligned} \tag{10}$$

Here the first summation $\sum_{i=1}^{2^d}$ is a large sum and so the calculation is computationally expensive. The number of comparisons required for a mixture model having J components modelling a data of dimensionality d is $2^d J^2$. Furthermore, we use ten-fold cross-validation in our experiments which further increases the complexity. This computation is feasible when the dimensionality of data is low ($d \ll n$), often less than 10.

We can enumerate all the possible states *present* in the data instead of enumerating all the possible states. The states absent in the data are improbable and the samples present in the dataset better approximate the KL divergence. Furthermore, using only the data samples in the data provides a data driven approach to approximating to KL divergence. Thus, $\sum_{i \in X^*}$ can approximate the summation $\sum_{i=1}^{2^d}$, where $\boldsymbol{X}$ denotes the dataset and $\boldsymbol{X}^* = \{x^* : x^* \in \boldsymbol{X}\}$ is a set of all the unique data samples present in the dataset. Here also i indexes all the unique samples in the dataset. Now, we can approximate (10) as:

$$KL_{\theta\beta} = \sum_{i \in x^*} \left[\left\{ \prod_{k=1}^{d} \left(\theta_k^{x_{ik}^*} (1-\theta_k)^{(1-x_{ik}^*)} \right) - \prod_{k=1}^{d} \left(\beta_k^{x_{ik}^*} (1-\beta_k)^{(1-x_{ik}^*)} \right) \right\} \cdot \sum_{k=1}^{d} \log \frac{\theta_k^{x_{ik}^*} (1-\theta_k)^{(1-x_{ik}^*)}}{\beta_k^{x_{ik}^*} (1-\beta_k)^{(1-x_{ik}^*)}} \right]. \tag{11}$$

In the Fig. 2, we empirically verify that there is no considerable loss of information when approximating KL divergence using the unique samples in the data. We sacrifice the accuracy of the KL divergence computation for gain in computational speed. We tabulate the gain in the computational efficiency in Table 1.

Again coming back to (8), we can define the probability distributions $P(x)$ and $Q(x)$ for the component distributions from a multivariate Bernoulli distribution in the following intervals:

$$P(x) \in [0,1] \text{ and } Q(x) \in [0,1] \Rightarrow \frac{P(x)}{Q(x)} \in [0,\infty] \text{ and } \log \frac{P(x)}{Q(x)} \in [-\infty,\infty] \tag{12}$$

Equation (12) shows that $\log \frac{P(x)}{Q(x)}$ in (8) has infinite range with the possibility of taking any values between $+\infty$ and $-\infty$. However, the probability terms $P(x)$ and $Q(x)$ are generally small because they are the product of numerous probability terms. For example, in chromosomal aberration dataset, it is always product of more than 8 (smallest dimensionality of data in chromosome 21) probability terms. Furthermore, we have a small background probability, ϵ>0, in our model such that both $P(x)$ and $Q(x)$ are never zeros.

The use of ϵ is also compensates for the left out state-space from possible 2^d samples. There will be no problem of normalization due to the addition of ϵ because we

Table 1 Difference in time requirement for the calculation of full KL divergence and our approximation

Chromosome (dimension)	Time in sec. for KL	
	Full	Approx
21 in Data 1 (8)	0.0992	0.0156
20 in Data 1 (10)	0.4863	0.0567
21 in Data 2 (14)	10.3447	0.0295
20 in Data 2 (20)	900.7118	0.0965

add ϵ to individual probabilities which is not greater than 1. Additionally, (8) describes the symmetric KL divergence and the choice of $P(x)$ and $Q(x)$ is arbitrary. Hence, multiplying (8) by log ratio only weighs the information in the difference measure of previous terms.

Figure 1 shows the scatter-plot of the minimum KL divergence obtained by our approximation against the full and accurate KL divergence between two random components. We experiment with one hundred, ten dimensional random models parameterized by six component distributions. The random models are mixture models initialized at random. However, the two components selected, based on the minimum KL divergence can mismatch between full and accurate KL divergence and our approximation. We can then merge two mistakenly selected components. Nevertheless, we compensate for such mismatches by retraining the mixture models after merging the mixture components.

It is important to note that we are primarily interested in determining the two closest component distributions in a mixture model. The exact minimum values between two component distributions in a mixture model or the true KL distances is of secondary concern. The chain of models we train and not the accuracy of approximation of the KL divergence determines the quality of our approximation. Figure 1 shows that our approximation of the KL divergence is less extreme but the minimum values are highly correlated with the full and accurate symmetrized KL divergence. Figure 1 also shows that the approximated KL divergence varies from 0.3 to 1 whereas the exact and accurate KL divergence varies from 4 to 16. This difference coupled with our good accuracy in selection of two components with the minimum KL divergence shows that our approximation ignores the terms that amplifies the difference but keeps the terms that contribute to the difference.

We used two different assumptions to approximate the full KL divergence: firstly, using only the samples present in the data and secondly, dropping the log term. We calculate the accuracy of our approximation compared to that of the full KL divergence to empirically verify that our approximation is suitable estimate of the full KL divergence. We need a mixture model and a dataset to run our algorithm.

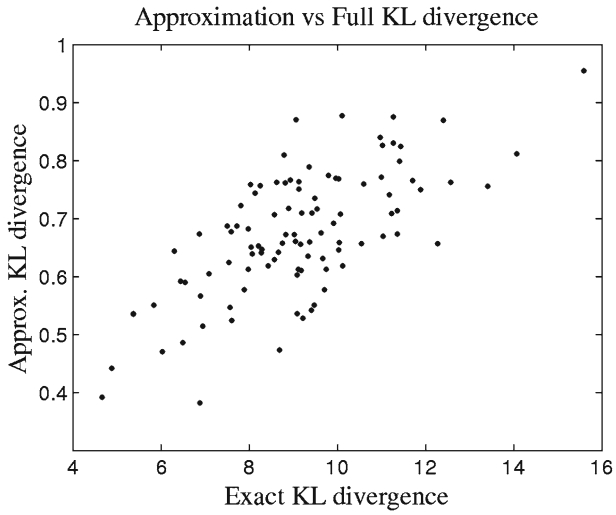

Fig. 1 Scatter-plot of minimum KL divergence values using our approximation of the KL divergence dropping the log term and also using the unique samples of data instead of the full binary state-space of the random variable x against the full and accurate KL divergence

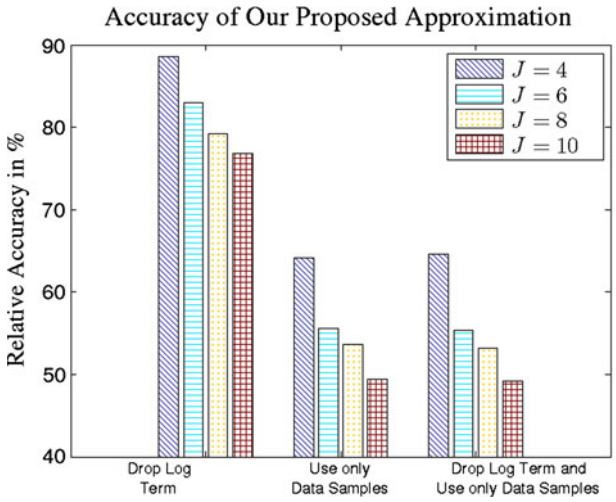

Fig. 2 The relative accuracy in the calculation of minimum KL divergence values using our approximation of the KL divergence dropping the log term and also using the unique samples of data instead of the full binary state-space of the random variable x

Thus, we first initialize four different mixture models with four, six, eight and ten components respectively. Secondly, we generate 5000 data points each from each mixture model. Thirdly, we use our algorithm to approximate two components in the mixture model.

Accuracy of calculation is two-fold because we approximate two components. In the experiments, either one of the two components matches or both the components match. Hence, we report both the accuracies showing that our approximation is similar to the full symmetric KL divergence. Figure 2 shows that the accuracy of matching of both the components is more than 50 % when the number of components is 10. Results of random matching is the combination of 10 components taken two at a time which is $1/\frac{n!}{k!(n-k)!} = 1/\frac{10!}{2!(10-2)!} \approx 2$ %. As the number of components gets higher, as expected, the accuracy gets lower.

Furthermore, Fig. 2 shows that most of the inaccuracy in the KL divergence approximation originates from using the data samples instead of the entire state-space of the random variable. However, using only the data–samples also contributes considerably to the speeding up of the algorithm. Finally, we can approximate the KL divergence between two component distributions omitting the log term and using only the samples present in the data as:

$$KL_{\theta\beta} = \sum_{i \in x^*} \left\{ \prod_{k=1}^{d} \left(\theta_k^{X_{ik}^*} (1-\theta_k)^{(1-x_{ik}^*)} \right) - \prod_{k=1}^{d} \left(\beta_k^{x_{ik}^*} (1-\beta_k)^{(1-x_{ik})} \right) \right\}. \tag{13}$$

3.2 Proposed fast progressive training of mixture models

Mixture models are widely used for clustering and our proposed algorithm in probabilistic modelling domain is similar to the hierarchical agglomerative clustering (Beeferman and Berger 2000). We merge the mixture components in the same manner as in the hierarchical agglomerative clustering. Additionally, we train all the parameters of the model subsequent to the merge operations. This is where our approach differs from the hierarchical agglomerative clustering. Furthermore, the hierarchical agglomerative clustering starts with the number of clusters equal to the

data points. In contrast, our algorithm starts with smaller values for the number of components such as 20 in our experiments, because the complexity of the mixture models is generally high.

Algorithm 1 Fast Progressive Training of Finite Mixture Models

Input: Dataset $\boldsymbol{X}$, and Maximum No. of Components J
Output: A series of Mixture models $\{\boldsymbol{\Theta_j}\}_{j=1}^{J}$ with different number of component distributions $j = 1, 2, \ldots J$

1: $\boldsymbol{\Theta}_J \leftarrow$ Best of 100 mixture models trained on data $\boldsymbol{X}$ having J components based on likelihood on $\boldsymbol{X}$
2: **for** j in J to 1 **do**
3: **if** $j! = J$ **then**
4: $\boldsymbol{\Theta}_j \leftarrow$ A trained mixture model on $\boldsymbol{X}$ using $\bar{\boldsymbol{\Theta}}_j$ as initialization
5: **end if**
6: **if** $j! = 1$ **then**
7: $(k^*, l^*) \leftarrow \underset{k,l}{\operatorname{argmin}} \ \boldsymbol{\mathcal{D}_{KL}}(p(x;\Theta_k)); p(x;\Theta_l))$

where $k, l \in (1 \ldots J); k \neq l$
8: $\bar{\boldsymbol{\Theta}}_{j-1} \leftarrow$ Mixture model where components π_{k^*}, π_{l^*} in $\boldsymbol{\Theta}_j$ are merged
9: **end if**
10: **end for**
11: **return** Series of mixture models $\{\boldsymbol{\Theta_j}\}_{j=1}^{J}$

Algorithm 1 shows the algorithmic flow of our proposed algorithm. In Algorithm 1, $\bar{\Theta}_j$ and Θ_j denote initialized and trained model having j components respectively. The algorithm consists of three main operations. Firstly, we calculate the KL divergence between different components in a mixture model to determine the mixture components having the minimum KL divergence (Step 7 in Algorithm 1). Secondly, we merge the mixture components with the minimum KL divergence (Step 8 in Algorithm 1). Finally, we retrain the mixture models (Step 4 in Algorithm 1). We can use the algorithm in conjunction with any model validation criterion such as cross–validation, MDL, AIC, and BIC. For example, in Adhikari and Hollmén (2012), we have listed an algorithm that uses cross–validation in conjunction with our strategy to select optimal number of mixture components in a mixture model.

3.3 Merging of mixture components

We select two components that have the minimum symmetric KL divergence in a mixture model (Ueda et al. 2000). We merge the selected components and their parameter values as in (14) and (15), respectively. Here, $\pi_{\text{klmin},1}$ and $\pi_{\text{klmin},2}$ are the two candidate mixing coefficients of the component distributions with the minimum KL divergence selected to merge. Similarly, π_{merged} is the mixing coefficient of the component distributions obtained after merging the two component distributions and $\pi_{\text{klmin},1}$ and $\pi_{\text{klmin},2}$.

$$\pi_{\text{merged}} = \pi_{\text{klmin},1} + \pi_{\text{klmin},2} \tag{14}$$

$$\Theta_{\text{merged}} = \frac{\pi_{\text{klmin},1} \times \Theta_{\text{klmin},1} + \pi_{\text{klmin},2} \times \Theta_{\text{klmin},2}}{\pi_{\text{klmin},1} + \pi_{\text{klmin},2}} \tag{15}$$

Similarly, we can merge the parameters according to the weight of the component distributions as in (15). Here, Θ_{merged} are the parameter vectors of the component π_{merged}, obtained by merging the two components in (14). Similarly, $\Theta_{\text{klmin},1}$ and $\Theta_{\text{klmin},2}$ are the parameter vectors of the two components $\pi_{\text{klmin},1}$ and $\pi_{\text{klmin},2}$ having the minimum KL divergence.

4 Experiments with fast progressive training of mixture models

We perform experiments with our proposed algorithm on one artificial dataset, one publicly available dataset, and two different real world chromosomal aberrations dataset. We use BernoulliMix, an open-source program package for finite mixture modelling of Multivariate Bernoulli distributions, to learn the parameters of mixture model using the EM algorithm.

We calculate the KL divergence between all the pairs of component distributions in the mixture model and select the one with the minimum KL divergence. We merge the selected components and their parameters as in (14) and (15). We obtain model with $(j-1)$ components by merging the two components in a mixture model having (j) components. This model with $(j-1)$ components will be an initialization model to train a final mixture model having $(j-1)$ components. This merging and retraining starts with 20 components and ends when the number of components becomes 1.

Figure 3 shows two adjoining mixture models that summarize the aberration patterns in cancer patients. Seven rows in the left denote seven components in the mixture model whereas 6 different rows in the right indicate 6 different components in the mixture model. The detached blocks in the left of each of the mixture models spanning one block each in each row visualizes the parameters of the component distributions while the adjoining blocks spanning eight blocks visualizes the parameters of that component distribution. Darker color denotes the higher values of the parameters and lighter color denotes lower values of the parameters. Adjacent to it on the right after the block arrow shows a mixture model having six components obtained after merging components 6 and 7 in the mixture model shown on the left panel. We train the mixture model in right panel to convergence. The correspondence between the components in the mixture models in the left and the right panels can be easily established. Components 1 to 5 in left panel panel corresponds to the components 1 to 5 in right panel. Combination of the components

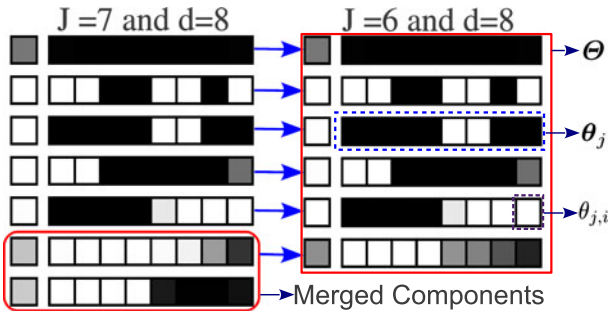

Fig. 3 Merging of components in the mixture model. Components 6 and 7 of the mixture model in the *left panel* of the figure have the minimum KL divergence of all the components. Hence, these two components are merged to form the single component in the mixture model in the *right panel* resulting in a mixture model having six components

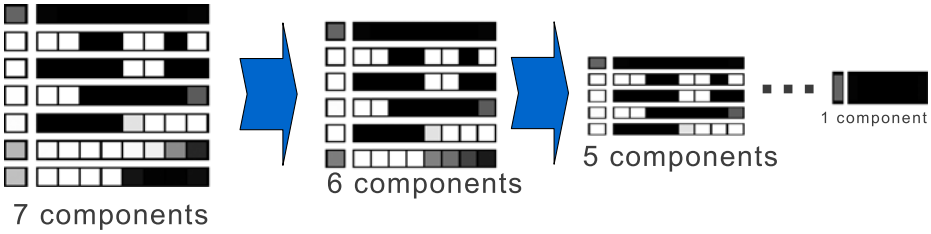

Fig. 4 Small snapshot of proposed algorithm where two components in a mixture models with 7 components are merged to generate a mixture model with 6 components and then another mixture model with five components progressively until we have a mixture model with a single component

6 and 7 in the model on the left panel results in component 6 in the model on the right of the left panel.

Our strategy for fast progressive training of mixture models is a search-based procedure that proceeds by going from complex to simple models, and is thus similar to the backward subset selection algorithm in feature selection literature (Kittler 1986). We have used a similar strategy in our sequential input selection algorithm SISAL in time–series prediction setting (Tikka and Hollmén 2008). However, here we select the component distributions unlike the data features (Tikka and Hollmén 2008).

We initially train the mixture model with a high number of mixture components, select two component distributions that are closest to each other, and merge them. This is progressively repeated until the number of components becomes 1. Furthermore, we restrict the maximum number of components to 20 because the highest dimensionality of data is 63 and mixture models with more than 20 components overfits the data. A mixture model with 20 components for data of dimensionality 63 consists of 1280 ($20 \times 63 + 20$) parameters. This big number of parameters is difficult to optimize with small size of data samples. Initially, we train 100 different models with 20 components via the EM algorithm using BernoulliMix program package. We select the best performing model of the 100 models for merging based on the likelihood in the data to minimize the problem of local optima of the EM algorithm.

Figure 4 shows the working of the algorithm. We merge the two components of a mixture model having 7 components to generate a mixture model with 6 components. We again merge the two components in a mixture model having 6 components to obtain a mixture model with 5 components. We repeat the procedure of merging of mixture components until we have a mixture model with a single mixture components. This progressive training results in a series of models of different complexities. We can use model validation techniques such as cross-validation with this strategy to select the optimal number of mixture components.

4.1 DNA copy number aberration dataset

We use two DNA copy number aberration datasets in the experiments. In the first data set, there are 393 different parts in the genome (data dimension, $d = 393$) (Hollmén and Tikka 2007; Myllykangas et al. 2008). Genome in the second data has 862 different parts (data dimension, $d = 862$) (Baudis 2007). We then transform the two available datasets to 0–1 matrix where rows denote the cancer patients, and

columns denote the chromosomal bands. $x_i = 1$ indicates that the chromosome band indexed by i for the cancer patient is aberrated whereas $x_i = 0$ indicates that the chromosome band is unaberrated.

Both the datasets used in the experiments have a limited number of samples; the first data consists of 4590 samples and the second data contains approximately 4000 samples. Thus, we perform chromosomewise mixture modelling because of the scarcity of data samples to constrain the complexity of mixture models. Mixture models trained with a small number of samples are susceptible to over–fitting and under-fitting. Unlike the mixture models of Bernoulli distributions, in Gaussian mixture model, higher number of mixture components sometimes helps in modelling the tails of the distributions.

When we divide the genome into chromosomes for chromosomewise analysis, each chromosome will have different number of chromosome bands, and different dimensionality. The reduced dimensionality will be considerably smaller than the original dimensionality of data. This decrease in dimensionality ameliorates the problem of curse of dimensionality (Donoho 2000). Chromosomewise mixture modelling will be computationally easier as the largest dimensionality is 63 (Chromosome 1) compared to the dimensionality of 862 for whole genome. Similarly, the smallest dimension is 8 (Chromosome 21) when we divide the genome with dimensionality 393 into different chromosomes. Furthermore, studying each chromosome separately can provide new insights into data; for example, chromosome specific patterns (Adhikari and Hollmén 2010a; Myllykangas et al. 2008) (Fig. 5).

We use a backward search based strategy to search for the optimal number of mixture components. We need to select the number of components based on their generalization performance on the future unseen data. Thus, we use the ten-fold cross-validation to train the model of each complexity where the number of components varies from 1 to 20. For each complexity, we obtain the initialization model by merging the two components of mixture model having one more mixture component.

Figure 6 shows that the likelihood smoothly decreases when the number of components decreases, i.e., increases with an increasing number of mixture components. It also shows that the increase in the likelihood with the increasing number of components is initially steep and then flattens out after a certain number

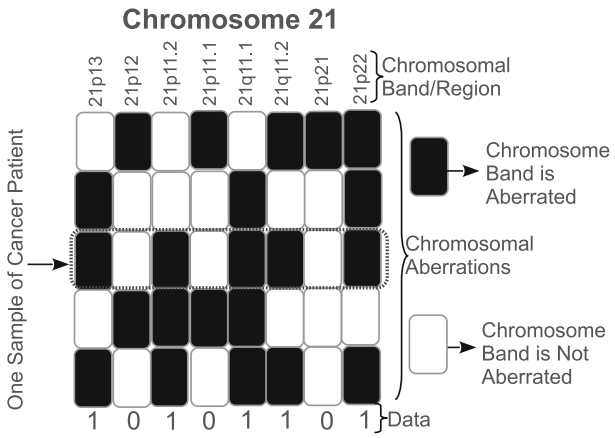

Fig. 5 Visualization of the dataset where rows are the cancer patients and the spatial co-ordinates on the X-axis are the chromosomal band. *Darker color* defines that the chromosomal band is aberrated and *lighter color* determines that chromosomal band is not aberrated. Figure shows only 5 samples in chromosome 21 for clarity

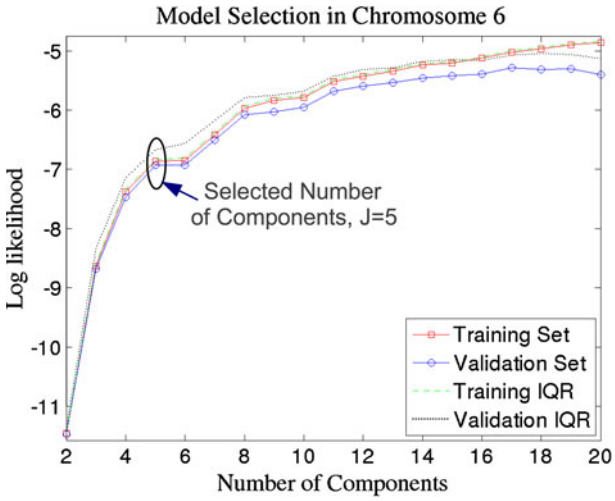

Fig. 6 An example of ten-fold cross-validation for model selection in Chromosome 6 in chromosomal aberrations data with dimensionality 393. The figure depicts averaged log-likelihood for training and validation sets. The interquartile range (IQR) for 10 different training and validation runs in ten-fold cross-validation setting have also been plotted. Here, number of components (J) selected is 5

of components. The figure also shows that the increase in the likelihood after the number of components is six is negligible considering the increase in the complexity of the model. Thus, we select 6 as the final number of components.

We also studied the changes in the likelihood values after merging the components and also after retraining the mixture model with the merged components. The results reported in the Fig. 7 shows that the log-likelihood decreases after merging of mixture components and increases after retraining it. The increase in likelihood obtained by retraining is unable to exactly compensate original value of likelihood of the mixture model having higher number of mixture components. However, figure shows that we achieve considerable improvement in log-likelihood values after retraining the mixture model obtained by merging of the mixture components. The improvement is greater in components 3 to 10 which are more likely to be the number

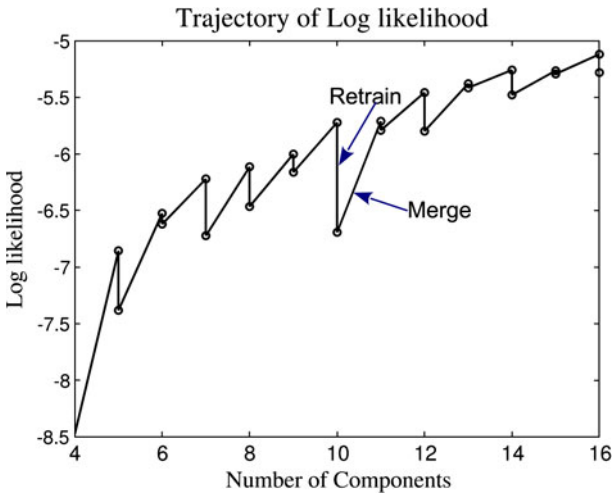

Fig. 7 An example of the trajectory of the log-likelihood in Chromosome 6 in chromosomal aberrations dataset with dimensionality 393. The figure shows log-likelihood on the whole data by the merged mixture model obtained by merging of the mixture components and the trained mixture model which is initialized using the merged mixture model. The model selection for the same data is shown in Fig. 6

of components selected for the data. Improvement decreases after the number of mixture components are greater than 10 showing that the models having more than 10 components may over-fit the data. Similarly, the improvement is unnoticeable when the number of components is less than 3 because the model may have under-fit. This repetitive training is one of the advantages of our proposed method over hierarchical agglomerative clustering.

4.2 Artificial dataset

Mixture models are generative models thus offering the facility to generate data samples from the model. This makes it easier to generate datasets with known number of components. In the experiments, we train the mixture model with six components and generate 3000 data points from the mixture model because our dataset of interest DNA Copy Number Aberration dataset contains similar number of samples. We also apply our algorithm of fast progressive training of mixture models on artificial data to determine if it correctly identifies six components that truly generated the data.

Left panel of Fig. 8 shows that our fast progressive training approach correctly identifies the six components present in the artificial data. Unlike the experiments with real–world datasets in Section 4.1 where we had no prior information about the number of components generating the data thus model selection is complicated for Chromosomal Aberration Datasets in Section 4.1. Nevertheless, we also experimented with adding noise (5 % and 10 %) to the artificial dataset. We flip the individual data elements from 1 to 0, and 0 to 1 of 5 % or 10 % of data elements to add 5 % or 10 % noise. The effect of noise is inconsiderable the algorithm from determining the true number of data generating components.

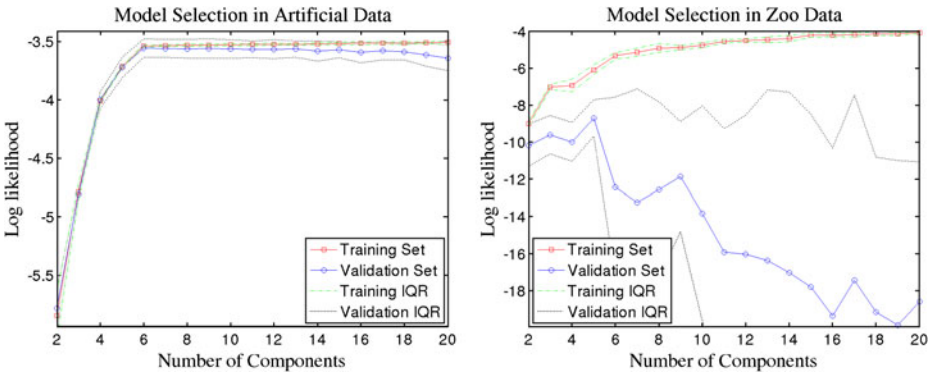

Fig. 8 *Left panel* shows ten-fold cross-validation for model selection in an artificial dataset. The figure depicts averaged log-likelihood for training and validation sets. The interquartile range (IQR) for 10 different training and validation runs in ten-fold cross-validation setting have also been plotted. Here, 6 components (J) generated the data and cross-validation results shows that we would have picked 6 components. Similarly, *right panel* shows ten-fold cross-validation for model selection in publicly available zoo dataset. The figure depicts averaged log-likelihood for training and validation sets. The interquartile range (IQR) for 10 different training and validation runs in ten-fold cross-validation setting have also been plotted. Here, the validation likelihood shows peaks at 3, 5 and 8 components. With background information on dataset, we would select 8 components

4.3 Publicly available zoo dataset

We also experimented our algorithm on a publicly available zoo dataset from UCI Machine Learning Repository (Bache and Lichman 2013). The dataset consists of 18 features. We ignore the first attribute out of the 18 features which is the unique animal name. Similarly, the last attribute is a categorical attribute that determines the cluster indices. The remaining 15 attributes of the 16 are binary attributes. The remaining one final attribute is a numerical attribute describing number of legs of the animal. We performed couple of experiments. Firstly, removing the numerical attribute completely. Secondly, changing the numerical attribute to six binary attribute as in Li (2005). The six attribute corresponds to the presence of 0, 2, 4, 5, 6, and 8 legs.

As in Li (2005), our algorithm identifies 5 clusters present in the zoo dataset. However, there are 7 clusters in the dataset. Right panel of the Fig. 8 shows high variation in likelihood as shown by the IQR curve when the number of components is 7. So, if we have a prior knowledge about the number of clusters, we can essentially select seven components because in different runs of the experiments maximum validation likelihood is when the number of components is 7. Furthermore, we perform experiments in five-fold cross-validation setting. The results in both the experiments were similar but sometimes inconsistent. One of the reasons for variation in results across different runs of cross–validation setting is because the number of samples 100 is considerably less to train the complexity of mixture models.

4.4 Improvement on previous model selection methods

We estimated the time required to compute our approximation of the KL divergence and also that of the full KL divergence to show the performance improvement with regards to the approximation of the KL divergence. The results reported in Table 1 shows that our approximation is considerably faster than the full KL divergence. Furthermore, it was computationally infeasible to calculate full KL divergence in the data of dimensionality greater than 20.

One of the major benefits of merging the mixture components is the faster convergence of the EM algorithm. The Fig. 3 shows that initialization model obtained by merging two components will be almost similar to the final model trained with six components. Therefore, number of iterations of the EM algorithm required to reach the convergence is considerably less. For example, in chromosome 21 in Dataset 1, on an average of 100 runs, a random model requires 47 iterations in the EM algorithm to converge while it requires only 1 iteration to converge from the initialized model. Similarly, on an average of 100 runs, it takes approximately 2.16 and 28.32 seconds to train a mixture model with the merged initialization and a random initialization respectively.

We used random initialization and repeated the experiments 50 times to ameliorate the local optima problem of the EM algorithm in our previous works in Tikka et al. (2007), Hollmén and Tikka (2007), Adhikari and Hollmén (2010a, b). Here we avoid 50 repeats because the EM algorithm is deterministic for the same data with the same initialization. For previous methods, in a ten-fold cross-validation setting with 20 components, we need to train $20 \times 10 \times 50 = 10000$ models whereas in this situation, we train only $20 \times 10 = 200$ different models. In other words, the

current proposed method makes one pass through the ten-fold cross validation setting whereas previous methods make 50 different passes through the ten-fold cross validation setting.

The advantage of our algorithm is production of similar models having different number of components by merging similar components. So, our algorithm compares similar models but with different number of components which makes the results more reliable as it helps us to find the accurate number of components. This avoids the situation where by random chance a model for accurate number of components gets stuck in local optima and with inaccurate number of components reaches the global optima. In that situation, we select inappropriate number of components. Although, our algorithm can get stuck in local optima, all models across all components gets stuck in similar local optima thus helping us select accurate number of components.

The additional overhead in the proposed method is the calculation of the KL divergence but as shown in Table 1, the approximations takes less than one-tenth of a second. Furthermore, yet another disadvantage of the merged initialization is that the EM algorithm can get stuck in a local optimum. However, we try to alleviate the problem by initially selecting the best of 100 models.

5 Summary and conclusions

In this paper, we proposed fast progressive training of mixture models to help model selection algorithms determine the number of mixture components by merging the mixture components. In the context of selecting which component distributions to merge, we proposed a fast, data driven approximation of the symmetrized KL divergence to calculate the similarity between two mixture components. Initially, we begin by selecting high number of mixture components and then progressively merge the similar components until the number of components is 1. We can use any model validation technique such as cross-validation, AIC, BIC, and MDL as a criterion for model selection in conjunction with our strategy. We experiment the proposed algorithm on two chromosomal aberration patterns data in cancer genomics showing that the our strategy produces plausible results. The proposed strategy is computationally efficient considering the well known pitfalls of methods using backward search strategy.

Acknowledgements Helsinki Doctoral Programme in Computer Science—Advanced Computing and Intelligent Systems (Hecse), and Finnish Center of Excellence for Algorithmic Data Analysis (ALGODAN) funds the current research.

References

Adhikari, P.R., & Hollmén, J. (2010a). Patterns from multi-resolution 0–1 data. In B. Goethals, N. Tatti, J. Vreeken (Eds.) *Proceedings of the ACM SIGKDD workshop on useful patterns (UP'10)* (pp. 8–12). ACM.

Adhikari, P.R., & Hollmén, J. (2010b). Preservation of statistically significant patterns in multiresolution 0–1 data. In T. Dijkstra, E. Tsivtsivadze, E. Marchiori, T. Heskes (Eds.) Pattern recognition in bioinformatics. *Lecture notes in computer science* (Vol. 6282, pp. 86–97). Berlin/Heidelberg: Springer.

Adhikari, P.R., & Hollmén, J. (2012). Fast progressive training of mixture models for model selection. In J.-G. Ganascia, P. Lenca, J.-M. Petit (Eds.) *Proceedings of fifteenth international conference on discovery science (DS 2012). LNAI* (Vol. 7569, pp. 194–208). Springer-Verlag.
Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.
Bache, K., & Lichman, M. (2013). *UCI machine learning repository*. University of California, Irvine, School of Information and Computer Science. http://archive.ics.uci.edu/ml.
Baudis, M. (2007). Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. *BMC Cancer, 7*, 226.
Beeferman, D., & Berger, A. (2000). Agglomerative clustering of a search engine query log. In *Proceedings of the ACM KDD '00, New York, USA* (pp. 407–416).
Blekas, K., & Lagaris, I.E. (2007). Split-merge incremental learning (SMILE) of mixture models. In *Proceedings of the ICANN'07* (pp. 291–300). Springer-Verlag.
Cai, H., Kulkarni, S.R., Verdú, S. (2006). Universal divergence estimation for finite-alphabet sources. *IEEE Transactions on Information Theory, 52*(8), 3456–3475.
Cover, T.M., & Thomas, J.A. (1991). *Elements of information theory*. New York: Wiley-Interscience.
Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal Of The Royal Statistical Society, Series B, 39*(1), 1–38.
Donoho, D.L. (2000) High-dimensional data analysis: the curses and blessings of dimensionality. Aide–Memoire of a lecture. In *AMS conference on math challenges of the 21st century*.
Everitt, B.S., & Hand, D.J. (1981). *Finite mixture distributions*. London, New York: Chapman and Hall.
Figueiredo, M.A.T, & Jain, A.K. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis Machicne Intelligence, 24*(3), 381–396.
Goldberger, J., Gordon, S., Greenspan, H. (2003). An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures. In *Proceedings of the ICCV '03, Washington DC, USA* (pp. 487–493).
Hershey, J.R., & Olsen, P.A. (2007). Approximating the Kullback Leibler divergence between Gaussian mixture models. In *IEEE. ICASSP 2007* (Vol. 4, pp. 317–320).
Juang, B.H., & Rabiner, L.R. (1985). A probabilistic distance measure for Hidden Markov models. *AT&T Technical Journal, 64*(2), 391–408.
Hollmén, J., & Tikka, J. (2007). Compact and understandable descriptions of mixture of Bernoulli distributions. In M.R. Berthold, J. Shawe-Taylor, N. Lavrač (Eds.) *Proceedings of the IDA 2007. LNCS* (Vol. 4723, pp. 1–12).
Kittler, J. (1986). *Feature selection and extraction. Handbook of pattern recognition and image processing.*. Academic Press.
Kullback, S., & Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*(1), 79–86.
Lee, Y.K., & Park, B.U. (2006). Estimation of Kullback–Leibler divergence by local likelihood. *Annals of the Institute of Statistical Mathematics, 58*, 327–340.
Leonenko, N., Pronzato, L., Savani, V. (2008). A class of Rényi information estimators for multidimensional densities. *Annals of Statistics, 36*(5), 2153–2182.
Li, T. (2005). A general model for clustering binary data. In *Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining, KDD '05* (pp. 188–197). ACM: New York.
Li, Y., & Li, L. (2009). A novel split and merge EM algorithm for gaussian mixture model. In *Fifth international conference on natural computation, 2009. ICNC '09* (Vol. 6, pp. 479–483).
Li, Y., & Li, L. (2009). A split and merge EM algorithm for color image segmentation. In *IEEE ICIS 2009* (Vol. 4, pp. 395–399).
Mclachlan, G.J., & Krishnan, T. (1996). *The EM algorithm and extensions* (1st ed.). Wiley-Interscience.
McLachlan, G.J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.
Myllykangas, S., Tikka, J., Böhling, T., Knuutila, S., Hollmén, J. (2008). Classification of human cancers based on DNA copy number amplification modeling. *BMC Medical Genomics, 1*(15), 1–18.
Perez-Cruz, F. (2008). Kullback–Leibler divergence estimation of continuous distributions. In *IEEE international symposium on information theory, ISIT 2008* (pp. 1666–1670).
Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing, 10*, 63–72.

Tikka, J., & Hollmén, J. (2008). A sequential input selection algorithm for long-term prediction of time series. *Neurocomputing, 71*(13–15), 2604–2615.

Tikka, J., Hollmén, J., Myllykangas, S. (2007). Mixture modeling of DNA copy number amplification patterns in cancer. In F. Sandoval, A. Prieto, J. Cabestany, M. Graña (Eds.) *Proceedings of the IWANN 2007. Lecture notes in computer science* (Vol. 4507, pp. 972–979). San Sebastián, Spain: Springer-Verlag.

Ueda, N., Nakano, R., Ghahramani, Z., Hinton, G.E. (2000). SMEM algorithm for mixture models. *Neural Computation, 12*(9), 2109–2128.

Wang, Q., Kulkarni, S.R., Verdú, S. (2005). Universal estimation of divergence for continuous distributions via data-dependent partitions. In *Proceedings international symposium on information theory, ISIT 2005* (pp. 152–156).

Windham, M.P., & Cutler, A. (1992). Information ratios for validating mixture analyses. *Journal of the American Statistical Association, 87*(420), 1188–1192.

Wolfe, J.H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research, 5*, 329–350.

Zhang, B., Zhang, C., Yi, X. (2004). Competitive EM algorithm for finite mixture models. *Pattern Recognition, 37*(1), 131–144.

Zhang, Z., Chen, C., Sun, J., Chan, K.L. (2003). EM algorithms for Gaussian mixtures with split-and-merge operation. *Pattern Recognition, 36*(9), 1973–1983.

Publication III

Prem Raj Adhikari, Jaakko Hollmén. Multiresolution Mixture Modeling using Merging of Mixture Components. In *Proceedings of Fourth Asian Conference on Machine Learning (ACML 2012)*, In Steven C.H. Hoi and Wray Buntine Editors, Volume 25 of Journal of Machine Learning Research—Proceedings Track, pages 17–32, November 4–6, 2012, Singapore, URL: http://jmlr.csail.mit.edu/proceedings/papers/v25/adhikari12.html, November 2012.

JMLR: Workshop and Conference Proceedings 25:17–32, 2012 Asian Conference on Machine Learning

Multiresolution Mixture Modeling using Merging of Mixture Components

Prem Raj Adhikari PREM.ADHIKARI@AALTO.FI
Jaakko Hollmén JAAKKO.HOLLMEN@AALTO.FI
Helsinki Institute for Information Technology and Department of Information and Computer Science
Aalto University School of Science, PO Box 15400, FI-00076 Aalto, Espoo, Finland

Editor: Steven C.H. Hoi and Wray Buntine

Abstract

Observing natural phenomena at several levels of detail results in multiresolution data. Extending models and algorithms to cope with multiresolution data is a prerequisite for wide-spread exploitation of the data represented in the multiple resolutions. Mixture models are widely used probabilistic models, however, the mixture models in their standard form can be used to analyze the data represented in a single resolution. In this paper, we propose a multiresolution mixture model based on merging of the mixture components across models represented in different resolutions. Result of such an analysis scenario is to have multiple mixture models, one mixture model for each resolution of data. Our proposed solution is based on the idea on the interaction between mixture models. More specifically, we repeatedly merge component distributions of mixture models across different resolutions. We experiment our proposed algorithm on the two real-world chromosomal aberration datasets represented in two different resolutions. Results show an improvement on the compared multiresolution settings.

Keywords: Multiresolution, Mixture Models, KL Divergence, 0-1 data

1. Introduction

Multiresolution data arise when an object or a phenomenon is described at several levels of detail. Multiresolution data are prevalent in several application domains such as image processing, hydrology, telecommunications, time series analysis, and astronomy (Willsky, 2002). The notion of multiresolution models is also related with the notion of multi-scale modelling (Cristini and Lowengrub, 2010; Ferreira and Lee, 2007) and wavelets (Mallat, 1989), thus widening the perspective of research on the multiresolution analysis.

Finite mixture models, or shortly mixture models, are probabilistic models widely used in the several analysis tasks such as clustering, density estimation, handling missing data, and modelling heterogeneity (Bishop, 2006; McLachlan and Peel, 2000). Mixture models are one of the most popular probabilistic modelling techniques due to their relative simplicity and flexibility to model more complex distributions based on a superposition of simple, parametric component distributions. In their standard form, however, the mixture models can be used to analyze the data in a single resolution.

A straightforward extension to handle data in the multiple resolutions is to model the data represented in different resolutions separately and to compare or combine the obtained results. Another option is to ignore one source of data and model only one of the available

data sets, which is represented in a single resolution. This effort results in less data and less representative data. The improvement on this method is to transform the data in different resolutions to a single resolution, integrate the data sets, and then apply the mixture models on the integrated data in the same resolution. This improves the performance over single analysis on the data in the different resolutions separately which has been shown in our previous work (Adhikari and Hollmén, 2010a,b).

All the previously mentioned solutions generate models in a single resolution. A more natural setting would be to have a separate model for each of the resolution, each of the models reflecting the properties of the data sets jointly. Our problem scenario is such that there are two or more datasets describing the same domain and which are expected to have the same distribution, but they have a different data dimensionality, or resolution. We learn the mixture models for each resolution so that we also model the interaction across different resolutions. Authors have tried to use the mixture models for the multiresolution data especially in the image processing domain (Wilson, 2000). However, the mixture of trees used in image compression and reconstruction in (Wilson, 2000) are not directly applicable in the other applications because the pyramid structure and the scale space in other applications are not as smooth as the one in the image processing. Furthermore, the data (features) in biology are often irregular thus necessitating a specialized approach to analyze the multiresolution data. In this paper, we propose a multiresolution mixture model based on the idea of merging of the mixture components across the different resolutions.

The concept of splitting and merging of the mixture components keeping their number fixed is used in (Ueda et al., 2000) and (Zhang et al., 2003) to ameliorate the problem of the local minima in the EM algorithm. Similarly, the authors in (Li and Li, 2009) and (Adhikari and Hollmén, 2012) use the split and merge strategy combined with a model selection criterion such as the Minimum Description Length (MDL) and the cross-validation (CV) to determine the optimal number of the mixture components in a mixture model (i.e. for model selection) varying the number of mixture components to search for the optimal number of the mixture components. The authors used the components within the same mixture model and only the two components are merged at any instant. In our proposed multiresolution mixture model, in contrast, we often merge more than the two mixture components from more than the two different mixture models. Similarly, in all of these studies, the authors do not consider the mixture models for the data in the multiple resolutions.

We train the mixture model separately in the different resolutions and merge the mixture components in the different resolutions thereby producing the mixture models in the multiple resolutions. We use the data driven fast approximation of the KL divergence to compute the similarity between the mixture components.

Our proposed algorithm of the multiresolution modelling is similar to clustering aggregation (Gionis et al., 2007), which for a given many clusterings generates a single clustering that agrees as much as possible with the initial, input clusterings. Here, we can view the mixture model in different resolutions as the different input clusterings and the multiresolution mixture model as the model that aggregates (agrees as much as possible) the information on the mixture models in the different resolutions. Our proposed algorithm produces the multiple mixture models in the different resolutions whereas the clustering aggregation produces only a single clustering result. Furthermore, clustering aggregation works in data-space where as our proposed algorithm works in model-space.

A single clustering algorithm can not be the best clustering algorithm for every situation and every dataset. Similar to the clustering aggregation, several clustering ensemble algorithms have been proposed with an aim to combine the different partitions obtained by the different clustering algorithms into a single clustering solution (Ghaemi et al., 2009; Vega-Pons and Ruiz-Shulcloper, 2011). Clustering ensemble improves the clustering solution with respect to the robustness, novelty, stability and confidence estimation, and parallelization and scalability (Topchy et al., 2004; Ghaemi et al., 2009). However, these clustering ensemble methods use the consensus functions such as relabeling, voting, mutual information, and co-association which are not directly usable neither in the mixture models nor in the multiple resolutions.

Topchy et al. (2004) used the mixture models for clustering ensembles but the results of multiple clustering methods are used as input features to the finite mixture models. The limitation of the method is that it is suitable for the data in a single resolution and the final models are also available only in a single resolution. Furthermore, the mixture models used in (Topchy et al., 2004) are not valid for patterns in the original space. However, we can reap the benefits of generative property of the mixture models only if the model is valid for the data in the original space. Additionally, the information while modelling the multiresolution phenomenon is best preserved while modelling in multiple resolutions simultaneously. Therefore, instead of modelling each resolution separately, we absorb the information contained in different resolutions in a single model and generate the models in multiple resolutions. We experiment the proposed algorithm on the chromosomal aberrations data in the multiple resolutions.

The rest of the paper is organized as follows. Section 2 briefly reviews the mixture models of the multivariate Bernoulli distributions. Section 3 discusses and derives the KL Divergence to compare the mixture components in the different mixture models in the multiple resolutions. Section 4 discusses the process of transforming the parameters of the mixture models of different dimensionality across different resolutions. Section 5 presents our proposed multiresolution mixture modelling algorithm. Section 6 contains the description of the experiments performed on the real-world dataset describing the chromosomal aberrations in two different resolutions. Section 7 summarizes the paper.

2. Mixture Models for 0-1 Data

Finite mixture models of multivariate Bernoulli distributions (Wolfe, 1970), composed as a sum of J component distributions are defined as

$$p(\boldsymbol{x}|\boldsymbol{\Theta}) = \sum_{j=1}^{\mathrm{J}} \pi_j \prod_{i=1}^{\mathrm{d}} \theta_{ji}^{x_i} (1-\theta_{ji})^{1-x_i}. \tag{1}$$

The data vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ consists of d elements, and $x_i \in \{0, 1\}$. The mixture proportions π_j satisfy the properties $\pi_j \geq 0, \forall j = 1, \ldots J$ and $\sum_{j=1}^{\mathrm{J}} \pi_j = 1$. The component distributions are parametrized with the Bernoulli parameters θ_{ji} containing parameters for each component distribution j and for each data vector element i. We can collect the mixture coefficients to a vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_j)$, and the parameters of the component distributions to a matrix $\boldsymbol{\Theta} = (\theta_{ji})$. The parameters of the mixture model of multivariate

Bernoulli distributions are now $\{J, \boldsymbol{\pi}, \boldsymbol{\Theta}\}$. Learning the mixture model requires estimation of the number of the mixture components, J, the mixing proportions, $\boldsymbol{\pi}$, and the parameters of the component distributions, $\boldsymbol{\Theta}$. If the number of mixture components J is assumed to be known, the EM algorithm (Dempster et al., 1977) can be used to learn the maximum likelihood estimates for the parameters of the mixture model.

Model selection in the context of mixture models refers to the problem of selecting an appropriate number of the component distributions. Several model selection algorithms have been proposed in the literature to learn the number of the mixture components in the mixture model (Smyth, 2000; Figueiredo and Jain, 2002). In our previous work, we have demonstrated the use of the model selection algorithms in the mixture models of the multivariate Bernoulli distributions (Tikka et al., 2007; Hollmén and Tikka, 2007; Adhikari and Hollmén, 2010a,b; Adhikari and Hollmén, 2012). In this paper, we are not concentrated on the problem of model selection but the proposed algorithm uses the trained models in the multiple resolutions and absorbs the information in multiple resolutions thereby generating models in the multiple resolutions.

3. Kullback-Leibler Divergence

Kullback-Leibler (KL) divergence is a non-symmetric measure of the difference between the two probability distributions (Kullback and Leibler, 1951; Kullback, 1959). Given two probability distributions P and Q, the KL divergence can be symmetrized by averaging the KL divergence from P to Q and from Q to P (Dagan et al., 1997). Mathematically, the symmetric KL divergence between the two probability distributions P and Q is given by:

$$\begin{aligned} \mathcal{D}_{KL}(P||Q) + \mathcal{D}_{KL}(Q||P) &= \sum_i P(i) log \frac{P(i)}{Q(i)} + \sum_i Q(i) log \frac{Q(i)}{P(i)} \\ &= \sum_i \left[\{P(i) - Q(i)\} log \frac{P(i)}{Q(i)} \right] \end{aligned} \tag{2}$$

where i indexes all the possible combinations of data elements.

The KL divergence between the two components in a mixture model to compare the two component distributions from a mixture model of the multivariate Bernoulli distributions has been derived in (Adhikari and Hollmén, 2012) as:

$$KL_{\alpha\beta} = \sum_{i \in X^*} \left\{ \prod_{k=1}^{d} \left(\alpha_k^{X^*_{ik}} (1 - \alpha_k)^{(1 - X^*_{ik})} \right) - \prod_{k=1}^{d} \left(\beta_k^{X^*_{ik}} (1 - \beta_k)^{(1 - X^*_{ik})} \right) \right\}. \tag{3}$$

Here, i indexes the unique samples denoted by X^* such that $X^* = \{x^* : x^* \in \overline{\underline{\boldsymbol{X}}}\}$ where $\overline{\underline{\boldsymbol{X}}}$ denotes the dataset. The two component distributions in a mixture model are denoted by α and β. Similarly, d denotes the dimensionality of data indexed by k. The Equation (3) constraints that both the component distributions should have the same dimensionality and should be indexed by the same dataset. We can extend this comparison to the multiresolution scenario under the simple and realistic assumption that the difference in the dimensionality contributes very less to the difference in the KL divergence as:

$$KL = \sum_{i \in X^*} \pi_\alpha \prod_{m=1}^{d} \left(\alpha_m^{X^*_{im}} (1-\alpha_m)^{(1-X^*_{im})} \right) - \sum_{i\prime \in Y^*} \pi_\beta \prod_{n=1}^{d'} \left(\beta_n^{Y^*_{i\prime n}} (1-\beta_n)^{(1-Y^*_{i\prime n})} \right) \quad (4)$$

Here, i and $i\prime$ indexes the unique samples in the two different datasets in two resolutions denoted by $\underline{\overline{\boldsymbol{X}}}$ and $\underline{\overline{\boldsymbol{Y}}}$ such that $X^* = \{x^* : x^* \in \underline{\overline{\boldsymbol{X}}}\}$ and $Y^* = \{y^* : y^* \in \underline{\overline{\boldsymbol{Y}}}\}$ are the set of all the unique data samples present in each dataset, respectively. Additionally, m and n indexes the dimensionality of datasets in the coarse and the fine resolution denoted by d and d', respectively. Here, α and β denote the component distributions in the two different mixture models in two different resolutions. Since Equation (4) approximates the symmetric KL divergence, the two terms in the equation can be interchanged. Furthermore, the Equation (4) is also suitable for cases when the number of data samples in the two different resolutions are different. In addition to the approximations, we weigh the KL divergence with their respective mixing proportions denoted by π_α and π_β in the Equation (4). When the KL divergence is weighted with the mixing proportions, it also considers the similarity of the mixing proportions which adds more suitability to comparing the component distributions from the different mixture models. Additionally, it is more desirable to merge the mixture components having the higher mixing proportions or having the lower mixing proportions as the mixing proportions also carries the information about similarity of the two mixture components in the context of the two different mixture models.

4. Sampling of Model Parameters

Merging the mixture components in the different models in the different resolutions is not straightforward because of the difference in the number of parameters (i.e. dimensionality d of the model parameters, $\boldsymbol{\Theta}$) of the component distributions. Therefore, we upsample the model parameters of the component distributions of the mixture models in the coarse resolution and downsample the parameters of the component distributions in the fine resolution to ensure that the dimensionality of the model parameters are the same. The concept of upsampling and downsampling is similar to that in the multiresolution data proposed in (Adhikari and Hollmén, 2010b) so that data in the different resolutions could be integrated. However, this paper proposes the upsampling and downsampling of the model parameters which allows seamless and simultaneous modelling of the multiresolution data. The model parameters are probabilities, not the 0-1 data as in (Adhikari and Hollmén, 2010b), therefore the upsampling and downsampling methods differs from the ones proposed in (Adhikari and Hollmén, 2010b). Furthermore, the sampling is performed in the model-space and not data-space thus providing simultaneous and seamless modelling of the multiresolution data.

4.1. Upsampling the model parameters

Upsampling transforms the model parameters of the component distributions from the coarse resolution to the fine resolution. In this case, one model parameter in the coarse resolution should produce multiple parameters in the fine resolution. We upsample the single model parameter by re-sampling the number of chromosomal regions required in

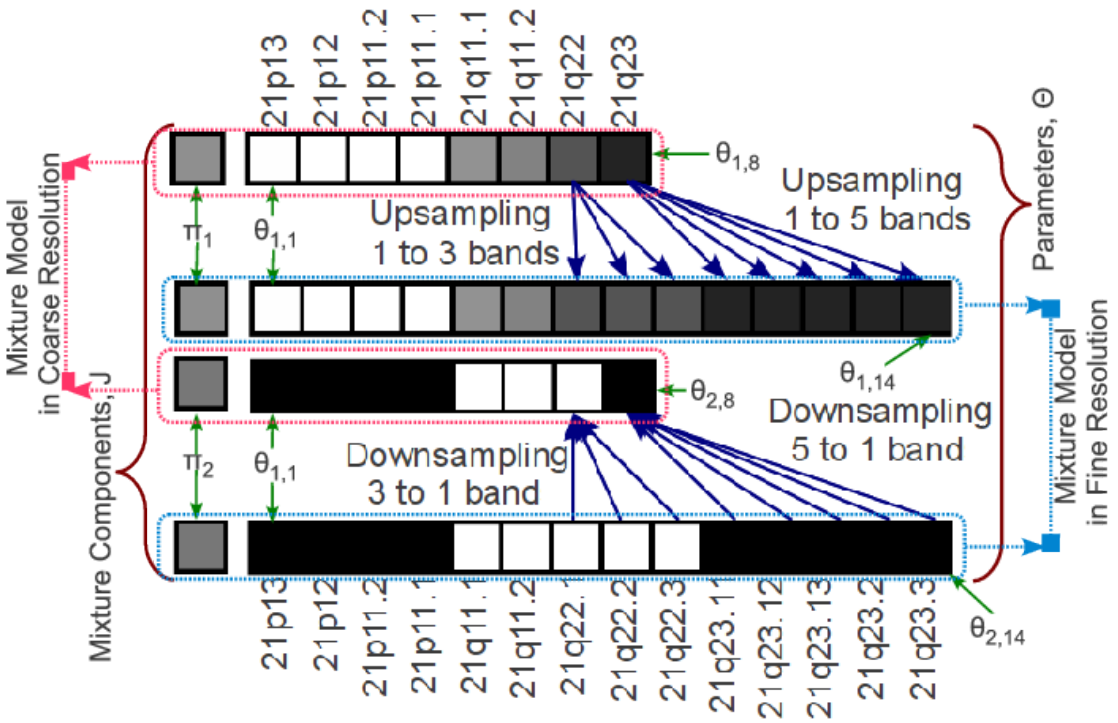

Figure 1: Illustration of the upsampling and the downsampling procedure for the model parameters in the two mixture models in the two resolutions. This is an example case in chromosome 21. The model parameters denote the regions of chromosome 21 and some of the chromosomal regions are unchanged across different resolutions as shown in (Shaffer and Tommerup, 2005). These unchanged chromosomal regions are not altered during sampling. However, other regions are upsampled from the coarse resolution and downsampled from the fine resolution according to the division of the chromosomal regions across different resolutions.

the fine resolution from a Normal distribution where the value of the model parameter in the coarse resolution is the mean and standard deviation is a small value (0.01 in our experiments). Since the model parameters are probabilities, we restrict the values of the model parameters θ between 0 and 1 $\{0 \leq \theta \leq 1\}$ by replacing the values that violate this constraint with the value of the model parameter in coarse resolution. Nevertheless, such deviations are rare because the value of standard deviation is small.

4.2. Downsampling the model parameters

Downsampling transforms the model parameters of the component distributions from the fine resolution to the coarse resolution combining the multiple parameters in the fine resolution to form a single parameter in the coarse resolution. We estimate the mean and standard deviation from the model parameters of the component distributions that are to be combined to downsample the model parameters. We then re-sample one model parame-

ter from a Normal distribution with estimated mean and the standard deviation. Similar to the upsampling, any value deviating from the probability range of $\{0 \leq \theta \leq 1\}$ is replaced with the mean of the model parameters in the fine resolution.

Figure 1 shows the upsampling and downsampling of the model parameters between the two different mixture models having two mixture components each. One detached block in the left in all the four rows visualizes the mixture components. The eight and fourteen adjoining blocks to the right shows the parameters of the component distributions. Darker color represents higher value for the mixture components and the model parameters while the lighter color denotes the smaller value for the mixture components and the model parameters. The mixing proportions are not changed while downsampling and upsampling. The solid arrows between the components across the different model parameters denotes the upsampling and downsampling procedures. The downward pointing arrows represent upsampling while the upward pointing arrows represent downsampling. The dotted arrows depict the two mixture models in two different resolutions.

4.3. Merging of Mixture Components

We select the components to be merged using the minimum weight bipartite matching (West, 1996) from the calculated symmetric KL divergence between the different components in the different mixture models. The updates are made in all the component distributions in all the mixture models initially by averaging all mixing proportions to be merged from the different mixture models as in the Equation (5) and subsequently to all the mixture components in a single mixture model during the normalization as in the Equation (6).

$$\pi_{merged} = \frac{\pi_{klmin,1} + \pi_{klmin,2} + \ldots + \pi_{klmin,n}}{n} \tag{5}$$

Equation (5) averages the selected mixing proportions in the different mixture models. Here, $\pi_{klmin,1}$, $\pi_{klmin,2}$, $\ldots$, $\pi_{klmin,n}$ indexes the mixture components having the minimum KL divergence merged together to form π_{merged} in the merged model. The update by the Equation (5) could violate the constraints in the mixture model such as the convex combination and the sum of probabilities as discussed in Section (2). Hence, the mixing proportions in each mixture model are finally normalized according to the Equation (6).

$$\pi_j = \frac{\pi_j}{\sum_{j=1}^{J} \pi_j} \tag{6}$$

where $j = 1 \ldots J$ indexes all the components in the mixture model.

$$\Theta_{merged} = \frac{\pi_{klmin,1} \times \Theta_{klmin,1} + \pi_{klmin,2} \times \Theta_{klmin,2} + \ldots + \pi_{klmin,n} \times \Theta_{klmin,n}}{\pi_{klmin,n} + \pi_{klmin,2} + \ldots + \pi_{klmin,n}} \tag{7}$$

The parameters can be merged according to the weight of the component distributions as given by the Equation 7. However, since the dimensionality of parameters of the component distributions are different, we upsample the parameters in the coarse resolution and downsample the parameters in the fine resolution. Secondly, we separately merge the mixture components in coarse resolution and the fine resolution using the Equation (7) producing merged model in each resolution.

5. Multiresolution Mixture Modelling Algorithm

Algorithm 1 provides the listings of the proposed algorithm to learn the multiresolution mixture model. The algorithm presents a simplified case of the multiresolution modelling which consists of the data in the two resolutions. The algorithm is scalable and expandable to N resolutions requiring $J(N-1)$ bipartite matching where J is the number of components. We do not need more than one comparison for any mixture model as we can move forward after selecting the similarity between the components in the first two mixture models. Given that a component a in the mixture model 1 is similar to a component b in the mixture model 2. If component c in mixture model 3 is similar to the component b in mixture model 2, then we can infer that the component a in the mixture model 1 is similar to the component c in the mixture model 3 without comparison. However, the parameters of the mixture model must be resampled (upsampled or downsampled) in all the different available resolutions.

Algorithm 1 Multiresolution Modelling using Merging of Mixture Components

Input: Two Datasets $\mathcal{D}_c$, $\mathcal{D}_f$, two mixture models M_c and M_f in coarse and fine resolution and a Threshold $\mathcal{T}_G$ for difference in KL divergence.

Output: Mixture Models mm_c and mm_f in coarse and fine resolution respectively that incorporates multiresolution data

klprev, indx $\leftarrow 0$

$\mathcal{T} \leftarrow \underset{k,l}{\text{argmax}} \;\; \mathcal{D}\{p(x \in \mathcal{D}_1; \Pi_k, \Theta_k); p(x \in \mathcal{D}_2; \Pi_l, \Theta_l)\}$

while $\mathcal{T}_G \leq \mathcal{T}$ **do**

  $merge_c$, $merge_f \leftarrow \emptyset$

  for i to min(Number of Components in m_c ($\mathcal{J}_c$) and m_f ($\mathcal{J}_f$)) **do**

    $(k^*, l^*) \leftarrow \underset{k,l}{\text{argmin}} \;\; \mathcal{D}\{p(x \in \mathcal{D}_1; \Pi_k, \Theta_k); p(x \in \mathcal{D}_2; \Pi_l, \Theta_l)\}$

    where $k \in (1 \ldots \mathcal{J}_c)$, $l \in (1 \ldots \mathcal{J}_f)$ $k \notin merged_c$, and $l \notin merged_f$

    $merge_c$, $merge_f \leftarrow$ insert k^*, l^*

    m_{c2f}, $m_{f2c} \leftarrow$ upsample(m_c), downsample(m_f)

    mm_c, $mm_f \leftarrow$ merge π_{k^*} and π_{l^*} in m_{c2f} and m_c, and in m_{f2c} and m_f

    $indx = indx + 1$

  end for

  if $mod(indx, 1000) == 0$ **then**

    mm_c, $mm_f \leftarrow$ Trained model on $\mathcal{D}_c$, $\mathcal{D}_f$ initialized using mm_c, mm_f

  end if

  $\mathcal{T} \leftarrow\mid klprev - \underset{k,l}{\text{argmin}} \mid$

  $klprev \leftarrow \underset{k,l}{\text{argmin}}$

end while

mm_c, $mm_f \leftarrow$ Trained model on $\mathcal{D}_c$, $\mathcal{D}_f$ initialized using mm_c, mm_f

return mm_c and mm_f

The input to the algorithm is the two datasets and the two trained mixture models in the coarse and the fine resolution and a threshold to stop the iterations for minimizing

the KL divergence. First, we calculate the symmetric KL Divergence between the different components of the mixture models in the different resolutions. Secondly, we upsample and downsample the model parameters so that they can be merged. We then match the components in the two different models using the minimum weight bipartite matching (West, 1996) and merge the components having the minimum KL Divergence. We retrain the mixture model via the EM algorithm initialized using the merged mixture model. We finally calculate the difference in the KL divergence between the two iterations. If the difference is less than the threshold, the algorithm ends by retraining the mixture models whereas if the change is not less than the threshold, we move on to the next iteration to minimize the KL divergence.

6. Experiments and Results

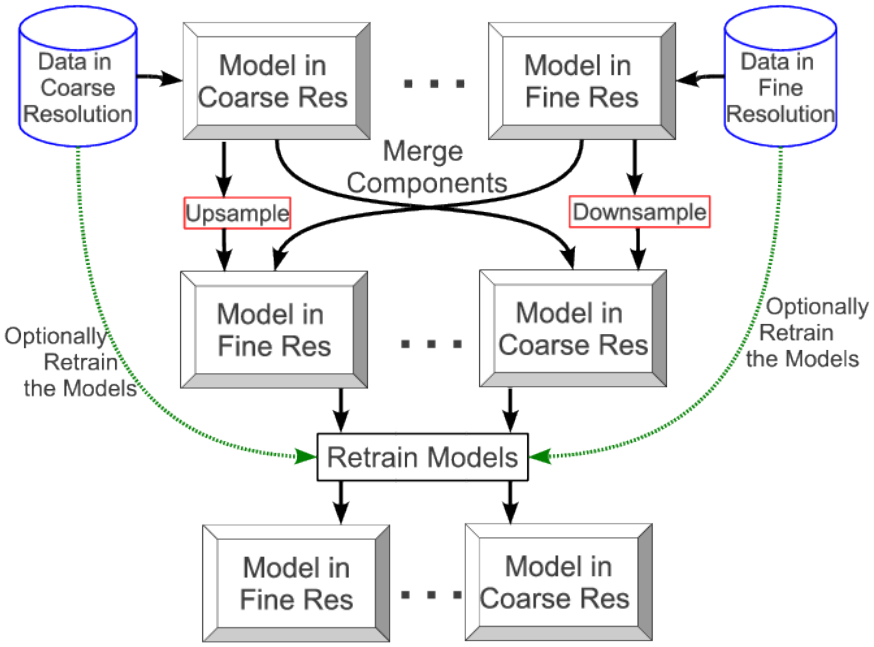

Figure 2: Multiresolution mixture modelling by merging of similar mixture components in different resolutions.

We experiment our proposed algorithm on the two chromosomal aberration patterns dataset in cancer genomics. One of the two datasets available from (Myllykangas et al., 2008; Hollmén and Tikka, 2007) was 393 dimensional (Coarse Resolution). In contrast, the other dataset available from (Baudis, 2007) was 862 dimensional (Fine Resolution). However, both the datasets explain the same phenomenon of chromosomal aberration and measure the similar chromosomal regions albeit in the different resolutions as explained in (Shaffer and Tommerup, 2005). The available datasets were converted to 0-1 matrix where each row denotes a cancer patient and each column denotes a region in the genome (a chromosome

band). Since the chromosomal aberrations data had small number of samples (≈4500) and high dimensionality, we experimented chromosome-wise on the data as in (Tikka et al., 2007). When the data for each chromosome is extracted from the genome, the different chromosomes will have the different dimensionality and some rows contain only zeros. Such rows with only zeros were removed because they contain no information with respect to the aberration pattern in the cancer patient in that chromosome.

Figure 2 summarizes the experimental procedure showing that the mixture models in two different resolutions are learned separately using the EM algorithm (Dempster et al., 1977). We then calculate the symmetric KL divergence between the different components in the two mixture models and match the components that have the minimum KL divergence using the minimum weight bipartite matching (West, 1996). The similar components are merged using the Equations (5) and (6). Similarly, the parameters of the component distributions are merged using the Equation (7). The merging of the mixture components are performed repeatedly until the changes in the KL divergence between the two mixture models in any two iterations is small (e.g. less than 10^{-3} in our experiments).

The estimated the time to compute our approximation of the KL divergence and also that of the full KL divergence to show the performance improvement gained by our approximation of the KL divergence in (Adhikari and Hollmén, 2012) shows that our approximation is considerably faster than the full KL divergence.

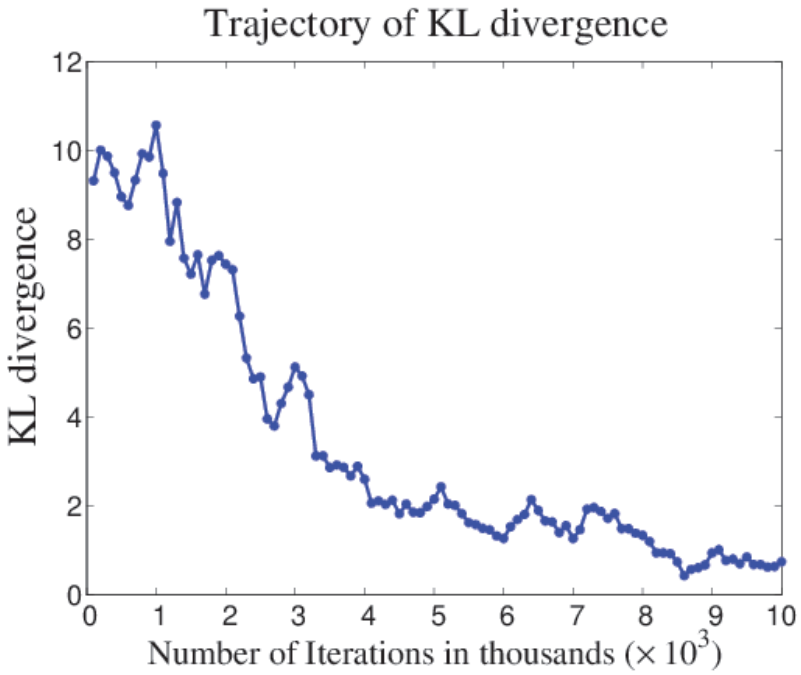

Figure 3: Changes in the KL divergence with increasing iterations.

Figure 3 shows the decrease in the KL divergence with the increasing iterations of minimizing the KL divergence. The Figure 3 is similar to convergence analysis as the KL divergence decreases with increasing number of iterations. Note in this case, the mixture model is not retrained. The decrease in the KL divergence is as expected not smooth

as some merging of components in the two resolutions will make differences in both the models at different resolutions in the different iterations. The KL divergence between the two models will approach to zero when both the models are similar to each other after repeated merging. However, it is not exactly zero in our case because of the upsampling and downsampling of the parameters makes the two models not exactly equal.

We used the upsampling and downsampling of the model parameters as discussed in Section (4) to merge mixture components in two different resolutions (having different dimensionality). The merged mixture model can be optionally trained on the combined data via the EM algorithm (Dempster et al., 1977) initialized using the merged mixture model. However, retraining the mixture model in each iteration of minimizing the KL divergence is computationally inefficient and the initialization model does not considerably vary in each iteration thus producing final model that is not different from the original model. Therefore, we can optionally retrain the mixture model in every thousandth iteration.

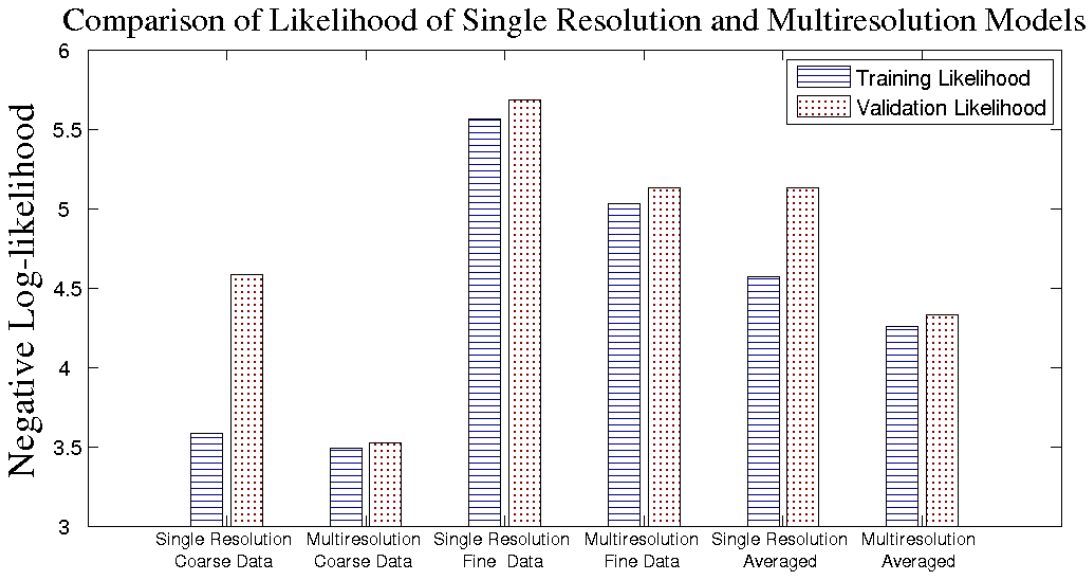

Figure 4: Likelihood of the multiresolution mixture model trained in a 10-fold cross-validation setting compared to the mixture model model in a single resolution. The result is an example case in the chromosome 17. Since the Y-axis shows the negative log-likelihood values, the shorter the bar the better the result.

We trained the multiresolution model and also the model in a single resolution in a ten-fold cross-validation setting via the EM algorithm. Here, the EM algorithm was initialized using the merged model in case of the multiresolution model and the single resolution models are initialized at random. Both the models were then trained to convergence via the EM algorithm. The multiresolution model is trained on the combined data obtained after integrating the data in two different resolution by transforming the data to the same resolutions as in Adhikari and Hollmén (2010b). However, the single resolution model were

trained with data in only one resolution ignoring the data in other resolution which is the current state-of-the-art.

The results in the Figure 4 show that the likelihood of the multiresolution models are better than that of the models trained on the data in the single resolution. Since the Y-axis in the figure shows the negative log-likelihood, the shorter the bar better the result. The Figure 4 shows three different cases of the likelihood: single resolution model on the coarse and the fine data, multiresolution model on the coarse and the fine data; and finally average of likelihood in the coarse and the fine data by multiresolution and single resolution model. The performance of the multiresolution model is markedly better in the coarse resolution and only slightly better in the fine resolution because the number of samples in the dataset is very small in the coarse resolution to add more information to the model in the fine resolution. Nevertheless, the average likelihood by the multiresolution shows noticeably improved performance of the multiresolution mixture models in both the resolutions.

Since the ten-fold cross-validation produces only ten values of the likelihoods which are very small to perform statistical significance testing on the result, we performed hundred-fold cross-validation producing 100 different training and validation likelihood values each. With the 100 different likelihood values, we performed the the two-tailed t-test to ascertain the statistical significance of our result. The results show that both the validation and the training likelihoods are statistically significant when the significance level, α, is 0.1. Training likelihoods in both the coarse and the fine resolution as well as the training likelihood in coarse resolution is statistically significant when the significance level, α, is 0.05. The validation likelihood in fine resolution is not significant when the significance level, α, is 0.05. The p-value of the validation likelihood in fine resolution is 0.09. This result in fine resolution can be attributed to the fact that the number of samples that have been added to the combined data by upsampling from the coarse data is small (342 samples have been added compared to 2716 samples in the fine resolution in the chromosome 17). which is considerably less to substantially improve the performance of the multiresolution model.

In order to show that our strategy of merging the mixture components positively facilitates the training of the mixture model via the EM algorithm, we calculated the iterations required by the EM algorithm to converge when initialized using the merged model. The left panel in the Figure 5 shows the number of iterations required as an average over five different runs of merging and retraining the models to converge to the final model via the EM algorithm. From the figure, we can see that the number of iterations required for the EM algorithm to converge decreases as we increase the iterations for minimizing the KL divergence. The decrease in the number of iterations shows that merging the mixture components moves the mixture model closer to the final model with regards to training via the EM algorithm. For each number of iterations, we restart the minimization of KL divergence such that the EM algorithm is used to train the final mixture model only once.

6.1. Illustration using models in same resolution

We experimented our algorithm on the two models in the same resolution to ascertain the improvements in the performance of the proposed multiresolution mixture modelling algorithm. As shown in in the right panel of the Figure 5, we selected two models such that the one fits the data poorly (solid line with the circles) while the another model fits

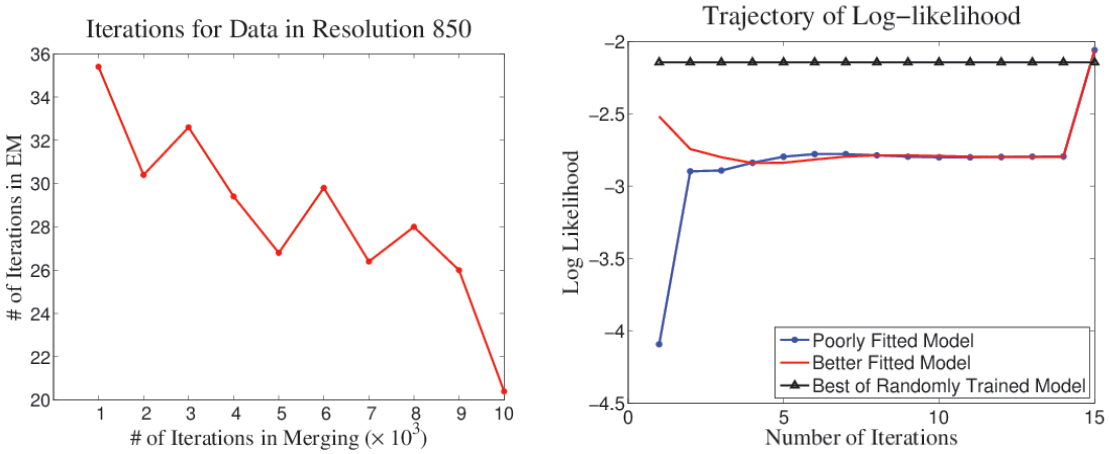

Figure 5: The left panel shows the number of iterations in the EM algorithm required to converge when the EM algorithm is initialized using the merged model after the minimization of KL is converged. The right panel shows the trajectory of the log likelihood values for two models in same resolution: a better fitted (solid line without the circles and the triangles) and a poorly fitted model (a solid line with the circles). The solid line with the triangles denotes the likelihood of the best of the 100 models trained on the combined data. The fifteenth iteration is the model retrained on the original data using the model after the 14th iteration as the initialized model. This is an example case in chromosome 21.

the data better (solid line without the circles and the triangles). We run our algorithm on the two models without upsampling and downsampling of model parameters as the number of model parameters are the same. The results obtained are visualized in the right panel of the Figure 5 showing that when the model converges, the likelihood on the combined data is better than the average likelihood of two models. However, the average likelihood is lesser than the best of randomly trained 100 models to convergence (solid line with triangles). Nevertheless, the likelihood obtained by merging the mixture components can be considered the validation likelihood as models are trained on the downsampled data and not on the combined data.

If we train the merged mixture model to convergence where the model obtained after the fourteenth iteration of merging is used to initialize the EM algorithm, we get the mixture model which produces better likelihood than the best of the randomly trained mixture model as shown in the fifteenth iteration in the right panel of the Figure 5. The improvement shows that our model is better than the best of the randomly trained model and also explains the importance of retraining the mixture model after every thousandth iteration. Furthermore, the fact that the merged mixture model producing better results experimentally verifies that our algorithm may be useful in avoiding local optima by making the little changes to the mixture models in the coarse and the fine resolution and also modelling the interactions across different resolutions. However, this result is neither mathematically proved and nor experimentally verified to work in every repeat of the experiment.

7. Summary and Conclusions

Multiresolution data arise when an object or a phenomenon is described at several levels of detail. In order to cope with the multiresolution data, standard data analysis methods need to be extended to include capabilities to model data in several resolutions simultaneously. In this paper, we proposed an algorithm for the multiresolution mixture modelling by merging the mixture components across the different resolutions. Given datasets in different resolutions from the same domain having similar distribution, our algorithm takes the models in different resolutions and repeatedly merges the mixture components minimizing the KL divergence thus producing a better mixture model. We performed experiments with our proposed algorithm on the two real-world chromosomal aberration datasets. The experiments show that our algorithm improves on the results of the competing multiresolution methods by involving the model interaction between the models in different resolutions.

Acknowledgments

This work has been funded by Helsinki Doctoral Programme in Computer Science - Advanced Computing and Intelligent Systems (Hecse), Helsinki Institute for Information Technology (HIIT), and Finnish Center of Excellence for Algorithmic Data Analysis Research (ALGODAN).

References

P. R. Adhikari and J. Hollmén. Preservation of statistically significant patterns in multiresolution 0-1 data. In Tjeerd Dijkstra, Evgeni Tsivtsivadze, Elena Marchiori, and Tom Heskes, editors, *Pattern Recognition in Bioinformatics*, volume 6282 of *Lecture Notes in Computer Science*, pages 86–97. Springer Berlin / Heidelberg, 2010a.

P. R. Adhikari and J. Hollmén. Patterns from multiresolution 0-1 data. In *Proceedings of the ACM SIGKDD Workshop on Useful Patterns*, UP '10, pages 8–16, New York, NY, USA, 2010b. ACM. ISBN 978-1-4503-0216-6.

P. R. Adhikari and J. Hollmén. Fast Progressive Training of Mixture Models for Model Selection. In J.-G. Ganascia, P. Lenca, and J.-M. Petit, editors, *Proceedings of Fifteenth International Conference on Discovery Science (DS 2012)*, volume 7569 of *Lecture Notes in Artificial Intelligence*, pages 145–156. Springer-Verlag, November 2012. ISBN 978-3-642-33491-7.

M. Baudis. Genomic imbalances in 5918 malignant epithelial tumors: An explorative meta-analysis of chromosomal CGH data. *BMC Cancer*, 7, 2007.

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

V. Cristini and J. Lowengrub. *Multiscale Modeling of Cancer: An Integrated Experimental and Mathematical Modeling Approach*. Cambridge University Press, 2010.

I. Dagan, L. Lee, and F. Pereira. Similarity-based methods for word sense disambiguation. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, EACL '97, pages 56–63, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38, 1977.

M. A. R. Ferreira and H. K. H. Lee. *Multiscale modeling: a Bayesian perspective*. Springer series in statistics. Springer-Verlag, 2007. ISBN 9780387708973.

M. A. T. Figueiredo and A. K. Jain. Unsupervised Learning of Finite Mixture Models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.

R. Ghaemi, N. Sulaiman, H. Ibrahim, and N. Mustapha. A Survey: Clustering Ensembles Techniques. *Proceedings of World Academy Of Science, Engineering and Technology*, 38: 644–653, 2009. ISSN 2070-3740.

A. Gionis, H. Mannila, and P. Tsaparas. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data*, 1(1), March 2007. ISSN 1556-4681.

J. Hollmén and J. Tikka. Compact and understandable descriptions of mixture of Bernoulli distributions. In M.R. Berthold, J. Shawe-Taylor, and N. Lavrač, editors, *Proceedings of the 7th International Symposium on Intelligent Data Analysis (IDA 2007)*, volume 4723 of *Lecture Notes in Computer Science*, pages 1–12, Ljubljana, Slovenia, September 2007. Springer-Verlag.

S. Kullback. *Information Theory and Statistics*. John Wiley and Sons, New York, 1959.

S. Kullback and R. A. Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Y. Li and L. Li. A Novel Split and Merge EM Algorithm for gaussian mixture model. In *ICNC '09. Fifth International Conference on Natural Computation, 2009.*, volume 6, pages 479–483, August 2009.

S. G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

G. J. McLachlan and D. Peel. *Finite mixture models*, volume 299 of *Probability and Statistics – Applied Probability and Statistics Section*. Wiley, New York, 2000.

S. Myllykangas, J. Tikka, T. Böhling, S. Knuutila, and J. Hollmén. Classification of human cancers based on DNA copy number amplification modeling. *BMC Medical Genomics*, 1 (15), May 2008.

L. G. Shaffer and N. Tommerup. *ISCN 2005: An International System for Human Cytogenetic Nomenclature(2005) Recommendations of the International Standing Committee on Human Cytogenetic Nomenclature*. Karger, 2005.

P. Smyth. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10:63–72, 2000. ISSN 0960-3174.

J. Tikka, J. Hollmén, and S. Myllykangas. Mixture Modeling of DNA copy number amplification patterns in cancer. In Francisco Sandoval, Alberto Prieto, Joan Cabestany, and Manuel Graña, editors, *Proceedings of the 9th International Work-Conference on Artificial Neural Networks (IWANN 2007)*, volume 4507 of *Lecture Notes in Computer Science*, pages 972–979, San Sebastián, Spain, 2007. Springer-Verlag.

A. P. Topchy, A. K. Jain, and W. F. Punch. A Mixture Model for Clustering Ensembles. In M. W. Berry, U. Dayal, C. Kamath, and D. B. Skillicorn, editors, *Proceedings of the Fourth SIAM International Conference on Data Mining (SDM)*. SIAM, 2004.

N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton. SMEM Algorithm for Mixture Models. *Neural Computation*, 12(9):2109–2128, 2000.

S. Vega-Pons and J. Ruiz-Shulcloper. A Survey of Clustering Ensemble Algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(3):337–372, 2011.

D. B. West. *Introduction to graph theory.* Prentice Hall, Second (Illustrated) edition, 1996.

A. S. Willsky. Multiresolution Markov models for signal and image processing. *Proceedings of the IEEE*, 90(8):1396–1458, August 2002. ISSN 0018-9219.

R. Wilson. MGMM: multiresolution Gaussian mixture models for computer vision. In *Proceedings of 15th International Conference on Pattern Recognition, 2000.*, volume 1, pages 212–215, 2000.

J. H. Wolfe. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5:329–350, 1970.

Z. Zhang, C. Chen, J. Sun, and K. L. Chan. EM algorithms for Gaussian mixtures with split-and-merge operation. *Pattern Recognition*, 36(9):1973–1983, September 2003.

Publication IV

Prem Raj Adhikari, Jaakko Hollmén. Mixture Models from Multiresolution 0–1 Data. In *Proceedings of Sixteenth International Conference on Discovery Science (DS 2013)*, Johannes Fürnkranz, Eyke Hüllermeier, and Tomoyuki Higuchi, Editors, Volume 8140 of Lecture Notes in Computer Science, Springer–Verlag, Berlin Heidelberg, pages 1–16, October 6–9, 2013, Singapore. DOI: 10.1007/978-3-642-40897-7_1 , October 2013.

Mixture Models from Multiresolution 0-1 Data

Prem Raj Adhikari and Jaakko Hollmén

Helsinki Institute for Information Technology (HIIT), and
Department of Information and Computer Science (ICS)
Aalto University School of Science,
PO Box 15400, FI-00076 Aalto, Espoo, Finland
{prem.adhikari,jaakko.hollmen}@aalto.fi

Abstract. Multiresolution data has received considerable research interest due to the practical usefulness in combining datasets in different resolutions into a single analysis. Most models and methods can only model a single data resolution, that is, vectors of the same dimensionality, at a time. This is also true for mixture models, the model of interest. In this paper, we propose a multiresolution mixture model capable of modeling data in multiple resolutions. Firstly, we define the multiresolution component distributions of mixture models from the domain ontology. We then learn the parameters of the component distributions in the Bayesian network framework. Secondly, we map the multiresolution data in a Bayesian network setting to a vector representation to learn the mixture coefficients and the parameters of the component distributions. We investigate our proposed algorithms on two data sets. A simulated data allows us to have full data observations in all resolutions. However, this is unrealistic in all practical applications. The second data consists of DNA aberrations data in two resolutions. The results with multiresolution models show improvement in modeling performance with regards to the likelihood over single resolution mixture models.

Keywords: Multiresolution data, Mixture Models, Bayesian Networks.

1 Introduction

A phenomenon or a data generating process measured in different levels of accuracy results in multiresolution data. This difference in accuracy arises because of improvement in measurement technology [1]. Newer generation technology measures finer units of data producing data in fine resolution. In contrast, older generation technology measures only the coarse units of data producing data in coarse resolution. Thus, accumulation of data over long duration of time results in multiresolution data. The availability of multiresolution data ranges across diverse application domains such as computer vision, signal processing, telecommunications, and biology [2].

The domain of scale-space theory [4], and wavelets [5] have close affinity with the domain of multiresolution modeling thus widening the scope of multiresolution modeling research. Furthermore, multiresolution data falls under one of the

J. Fürnkranz, E. Hüllermeier, and T. Higuchi (Eds.): DS 2013, LNAI 8140, pp. 1–16, 2013.

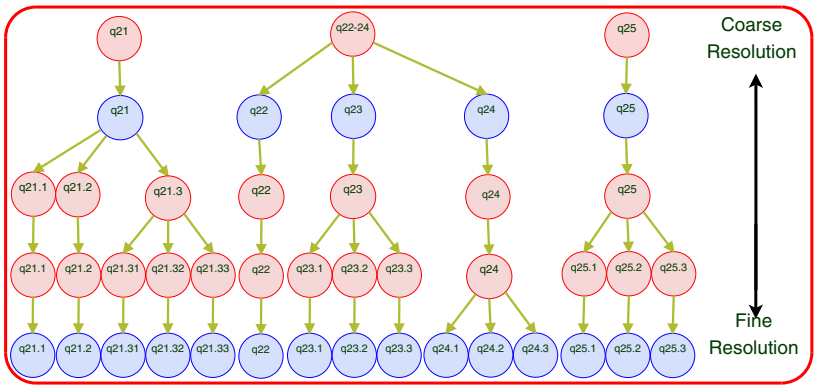

Fig. 1. A typical dichotomy of universals and particulars in giving rise to multiresolution data. Figure shows a part of chromosome–17 in five different resolutions of nomenclature as defined by ISCN [3].

essential ontological dichotomies of universals and particulars [6]. For example in cytogenetics, an application area of interest, International System for Human Cytogenetic Nomenclature (ISCN) has a standardized nomenclature for the parts of the genome. It has defined five different resolutions of the chromosome band: 300, 400, 550, 700, and 850 [3]. In other words, there are 862, and 311 regions in a genome in resolution 850 (fine resolution), and 300 (coarse resolution), respectively. Figure 1 shows an example of multiresolution data resulting from the ISCN nomenclature which is also our application area of interest. Figure 1 shows a part of chromosome–17 in five different resolutions forming a tree structure among different chromosome bands. Here, the same part of genome measured in different levels of detail generating multiresolution data.

Finite Mixture Models are semi-parametric probability density functions represented as the weighted sum of component densities of chosen probability distributions such as Gaussian, Bernoulli, or Poisson [7,8]. Mixture models have found wide spectrum of uses such as clustering [9], density estimation [10], modeling heterogeneity [11], handling missing data [12], and model averaging. They are versatile because of their suitability for any choice of data distribution, either discrete or continuous, and flexibility in the choice of component distributions [8]. However, mixture models in their basic form only operate on single data resolution, and is unable to model multiresolution data. The only mixture modeling solution to multiresolution data are to model the different resolutions separately and at best compare the findings. Cancer is not a single disease but a heterogeneous collection of several diseases [13]. Therefore, we use mixture models to model cancer patients discussed in Section 4.2 because mixture models are well known for their ability to model heterogeneity.

In our previous work, we transform the multiresolution data to a single resolution using different deterministic transformation methods, and model the resulting single resolution data [14]. Results in [14] shows improvement in the performance of mixture models through multiresolution analysis compared to

single resolution analysis. We also proposed a multiresolution mixture model based on merging of mixture components across different resolutions in [15]. The improvement in [15] is that the models assimilate the information contained in other data resolutions. Furthermore, transformations here are in the model domain unlike in the data domain as in [14]. In all scenarios above, the mixture models are generated in a single resolution and directly unusable in multiresolution scenarios without modification.

In the past, research has considered formulating the multiresolution mixture model in different application areas. For instance, the multiresolution Gaussian mixture model in [16] approximates the probability density, and adapts to smooth motions. The design of the model is for a specific choice of data distribution, and generates trees of decreasing variance, and consequently a tree of Gaussians. Unlike an actual multiresolution model, this essentially models single resolution data using a tree from the same data for multiple Gaussians on different scales with different variance. Furthermore, difference in the pyramid structure present in other domains limit its general applicability.

Authors have increased the efficiency and robustness of learning mixture models using multiresolution kd-trees [9,17]. Furthermore, authors in [10] have proposed the a mixture of tree distributions in a maximum likelihood framework. Similarly, authors in [18,19] use multiresolution binary trees to learn discrete probability distribution. However, it is impossible to represent all multiresolution data as kd-trees which are binary in nature. In addition, the focus in [18,19] is in modeling single resolution data. Additionally, multiresolution trees are also used in object recognition [20], and as binary space portioning trees [21]. However, the authors in [20,21] use them in the context of recursive neural networks, and geometric representations for information visualization, respectively.

In this paper, we propose a multiresolution mixture model whose components are Bayesian networks denoting the hierarchical structure present in multiresolution data. We learn the parameters of each component distribution in a Bayesian network framework. Component distributions in the form of Bayesian network is useful also to impute the missing data resolutions considering them as missing values. Finally, we transform the multiresolution data in the Bayesian network representation to a single data in vector form to learn the mixing coefficients and the parameters of the mixture model in a maximum likelihood framework using the EM algorithm.

2 Bayesian Networks of Multiresolution Data

Bayesian networks bring the disciplines of graph theory and probability together to elegantly represent complex real-world phenomena dealing with uncertainty [22,23]. A Bayesian network consists of nodes or vertices that encode information about the system in the form of probability distributions, and links or arcs or edges that denote the interconnections or interactions between nodes in the form of conditional independence [22]. It analytically represents a joint distribution over a large number of variables. Furthermore, it treats learning

and inference simultaneously, seamlessly merges unsupervised and supervised learning, and also provides efficient methods for handling missing data [23].

2.1 Component Distributions of Multiresolution Hierarchy as Bayesian Networks

A Bayesian network is a directed acyclic graph (DAG) that describes a joint distribution over the set of random variables $X_1, \ldots, X_d$ such that

$$P(X_1, \ldots, X_d) = \prod_{i=1}^{\mathrm{d}} P(X_i \mid \text{parents}(X_i)), \tag{1}$$

where parents(X_i) are the set of vertices from which there is an edge to X_i. Figure 2 shows an example of a Bayesian network of six random variables A, B, C, D, E, and F. The six vertices represent the six random variables, and the five directed edges represent the conditional dependencies (independence). We can define conditional independence in a Bayesian network as: A variable is conditionally independent of all the variables in the network given its Markov blanket. The Markov blanket of a variable is the set of its parents, its children, and the other parents of its children.

Data in multiple resolutions share a commonality because they measure the same phenomenon. A single feature in the coarse resolution corresponds to one or more features in the fine resolution. We can exploit this information from the application area to determine the relationships between data resolutions

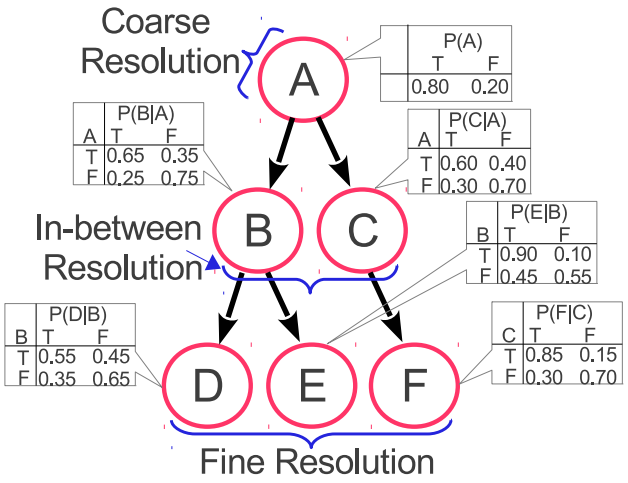

Fig. 2. Network representation of the multiresolution data where ancestors denote data in coarse resolution and leaves denote data in fine resolution. The network a simple Bayesian network of six random variables with five edges along with the associated probability tables.

and consequently, the structure of the Bayesian network. The data features in the coarse resolution form the root and branches near the root of the network. Similarly, the data features in the fine resolution form the branches towards the leaves and the leaves of the tree. Additionally, we can assume that the directed arrows originate from the features in the coarse resolution for computational efficiency.

Each vertex of a Bayesian network bears a corresponding conditional probability distribution (CPD). The CPD specifies that a child takes a certain value with a probability depending the value of its parents [22]. In the figure, for example the variables D, and E are conditionally independent given B. We can simplify the joint probability distribution of A, B, C, D, E, and F using the conditional independences in Figure 2 as:

$$\begin{aligned} P(A,B,C,D,E,F) &= P(A|B,C,D,E,F)P(B|A,C,D,E,F)\ldots \\ &\quad P(C|A,B,D,E,F)P(D|A,B,C,E,F)\ldots \\ &\quad P(E|A,B,C,D,F)P(F|A,B,C,D,E) \\ &= P(A)P(B|A)P(C|A)P(D|B)P(E|B)P(F|C) \end{aligned} \tag{2}$$

The CPD of a discrete variable is represented in a table as shown in Figure 2. It enumerates each possible set of values for the variable and its parents. Algorithms based on maximum likelihood (MLE) and maximum a posteriori estimates (MAP) can learn the parameters of the Bayesian network with a known structure [24]. In our application, the structure of Bayesian networks comes from the domain knowledge. The depth of the Bayesian network depends on the number of resolutions in multiresolution data. Learning a Bayesian network of known structure involves determining the CPD of the variables in the network. We learn the CPD of the variables using the Maximum Likelihood Estimate (MLE) [22].

2.2 Missing Resolutions in Multiresolution Data

Missing data has received considerable research interest because of their abundant occurrence in many domains [12,25]. The problem of missing data escalates when some resolutions (entire data) in a multiresolution setting are missing. Therefore, when the values are missing in multiresolution analysis, one or more resolutions (entire data) will be missing. This is unlike the typical missing data problems where small number of variables in some samples will be missing. For example, data in a coarse resolution can be missing while data in other resolutions are available. Bayesian networks have a seamless ability to handle missing data [12]. Therefore, learning the Bayesian networks also helps to generate the data in missing resolutions because Bayesian networks are generative models.

We can impute the missing values using marginal inference in Bayesian Networks. Marginal inference is the process of computing the distribution of a subset of variables conditioned on another subset [22]. We can calculate the marginal inference for a joint distribution P(A,B,C) given the evidence B=true as:

$P(A \mid B = true) \propto \sum_{C} P(A, B = true, C).$

Authors have proposed algorithms such as variable elimination, and sum product algorithm to compute marginal inference [22]. We draw samples under the given evidence from consistent junction trees using the BRMLToolbox [22].

3 Multiresolution Mixture Model of Multivariate Bernoulli Distributions

Mixture Models are semi–parametric latent variable models that models a statistical distribution by a weighted sum of parametric distributions [7,8,22]. They are flexible for accurately fitting any complex data distribution for a suitable

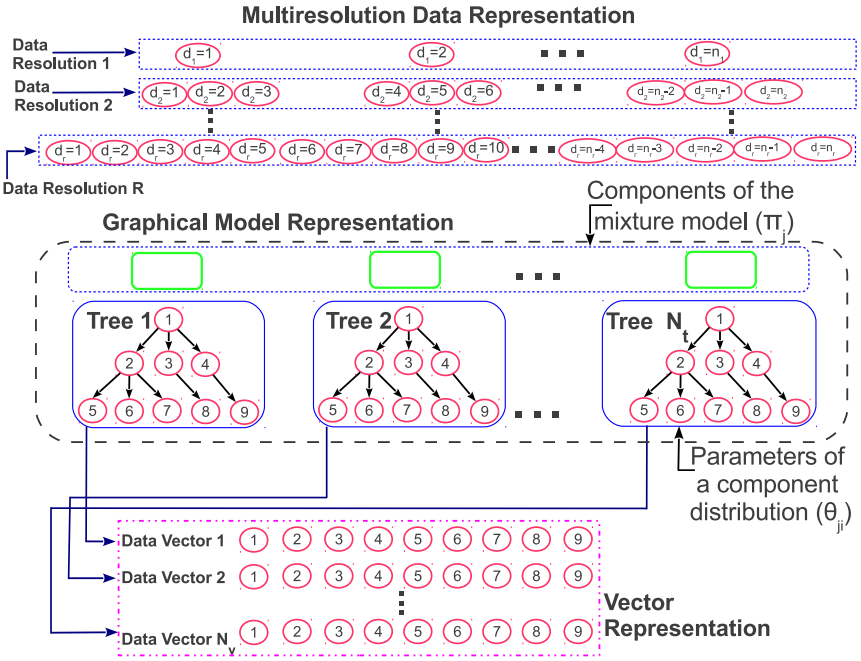

Fig. 3. Top panel shows the general representation of multiresolution data. There are 'r' data resolutions of different dimensionality. The multiresolution data representation is then transformed to a Bayesian network representation which in turn is mapped to a vector of single resolution data. The Bayesian network representation in the middle panel (within the dashed rectangle) also depicts a mixture model having the Bayesian networks as the components for data in multiple resolutions. The three solid rectangles on the top represent different mixture coefficients. Similarly, the three network of nodes denote the three component distributions where each vertex defines a parameter of the component distribution. The numbers inside the nodes denote the position of the variable in the vector representation with regards to dimensionality. The dash–dotted rectangle in the bottom of the figure shows the vector representation of the data derived from the Bayesian network representation. In the figure, N_t, and N_v denote the number of networks abreast adjacently in the multiresolution data, and the number of samples in multiresolution data mapped to a vector representation, respectively.

choice of data distribution, and a good enough number of mixture components. Expectation Maximization (EM) algorithm helps to learn the maximum likelihood parameters of the mixture model [26]. The EM algorithm requires prior knowledge of the number of components in the mixture model. Model selection is the process of determining the number of components in the mixture model [8]. In our previous work, we have tried to solve the problem of model selection in mixture models [11,27,28]. We learn mixture model of different complexities in a cross-validation setting and select a model with the number of components that gives the best generalization performance.

A multiresolution data is a collection of different component distributions as shown in the middle panel (within the dashed rectangle) of Figure 3. The multiresolution components in the middle panel (within the dashed rectangle) of Figure 3 encode the relationships between different resolutions of multiresolution data. The structure of the component distribution comes from the domain knowledge. Thus, the problem with regard to model selection in multiresolution data culminates to determining the optimal number of such component distributions present in the data. Similarly, learning the parameters of the component distributions involves learning the parameters of those networks.

In the general framework for the EM algorithm, we can assign only a single probability value to a node in the mixture model [26]. However, each variable in Bayesian network consists of minimum of two probability values denoting the CPD of the nodes. Therefore, in this contribution, we map the Bayesian network to vector representation to learn a multiresolution mixture model of Bayesian networks. This simple and intuitive solution proposed in this contribution transforms Bayesian networks to vectors with increasing dimensionality representing increasing depth of the Bayesian network. The first element of the vector will be the root node in the first generation i.e. coarsest resolution. Similarly, the last element of the vector will be leaf node of the last generation i.e. finest resolution arranged from left to right as shown in the bottom panel (within the dashed dotted rectangle) of Figure 3. In the middle and the bottom panel of Figure 3, the number inside the vertices denote the relative position of the variable in the vector representation with regards to dimensionality. Multiresolution mixture components transformed to a vector representation will have the same dimensionality because the structure of component distributions are identical with one another. However, component distributions have different parameters.

Vector representation of Bayesian networks eases modeling multiresolution data in one resolution. Furthermore, it increases the number of data samples. The number of samples in the vector form, N_v will be $N_v = N \times N_t$. Here, N is the number of samples of data in each resolution. Similarly, N_t is the number of Bayesian networks present in the data along the dimension corresponding to one sample. Furthermore, the data dimensionality will be considerably reduced as the depth of the Bayesian network is generally small. Additionally, Bayesian networks provide the sparsity [24], making data dimensionality in the vector representation d_v is smaller than that of the finest resolution of the multiresolution data. Furthermore, the vector representation has larger number of data samples

than the original Bayesian networks representation, $d_v < \max(d_r) \ll N_t \ll N_v$. Here, d_r is the dimensionality of data in different resolutions. An increase in the number of data samples and a reduction in dimensionality facilitates the learning of mixture models because they require a large number of samples to accommodate the increasing dimensionality of the data.

We can describe the mixture model of multivariate Bernoulli distributions for a 0-1 data [7] as:

$$p(\mathbf{x} \mid \Theta) = \sum_{j=1}^{\mathrm{J}} \pi_j \prod_{i=1}^{\mathrm{d}} \theta_{ji}^{x_i} (1 - \theta_{ji})^{1-x_i}. \tag{3}$$

Here, $\boldsymbol{\Theta} = \{J, \{\pi_j, \theta_j\}_{j=1}^{J}\}$ denotes the parameters of the mixture model of multivariate Bernoulli distributions. Here, π_j denotes the mixing proportions which sum to 1. Similarly, θ_{ji} is the probability that a random variable of the j^{th} component in the i^{th} dimension will take the value of 1. In multiresolution scenario, i differs for each resolution. Therefore, we have to model different resolutions with different models. We can formulate Equation 3 with respect to log likelihood to learn the mixture model using the EM algorithm in a maximum likelihood framework [8] as:

$$\mathcal{L}(\Theta) = \sum_{n=1}^{\mathrm{N}} \log\ P(x_n \mid \Theta) = \sum_{n=1}^{\mathrm{N}} \log \left[\sum_{j=1}^{\mathrm{J}} \pi_j \prod_{i=1}^{\mathrm{d}} \theta_{ji}^{x_{ni}} (1 - \theta_{ji})^{1-x_{ni}} \right]. \tag{4}$$

Given the number of mixture components, J, parameterized by $\Theta = \{\pi_j, \theta_j\}$, the EM algorithm can learn the mixture model that maximizes the likelihood in Equation 4. Model selection using cross-validated log likelihood can determine the number of mixture components, J [11,27,28].

4 Experimental Data

We experiment with the proposed methodology on two multiresolution data: an artificial data, and a chromosomal aberrations data, both in three resolutions.

4.1 Artificial Multiresolution 0-1 Dataset

In some application areas the relationships between different resolutions in a multiresolution setting are well known [6]. We can exploit such knowledge to artificially generate realistic multiresolution data. We initially fixed the structure of a Bayesian network to the the components of the mixture model shown in the middle panel (within the dashed rectangle) of Figure 3. Five such Bayesian networks were abreast along the dimensionality of the data. The dimensionality of data in three different resolutions are 5, 15, and 25 respectively. Firstly, we generate the data in the finest resolution i.e. having dimensionality of 25.

We fix two parameters to sample the data of given dimensionality: X is uniformly distributed in the range $[0, 1]$, and l is normally distributed with mean 0.1 and standard deviation 0.04.

$$X \sim U[0,1] \times d \text{ and } l \sim N\left(\mu = 0.1, \sigma^2 = 0.04\right) \times d$$

where d denotes the data dimension.

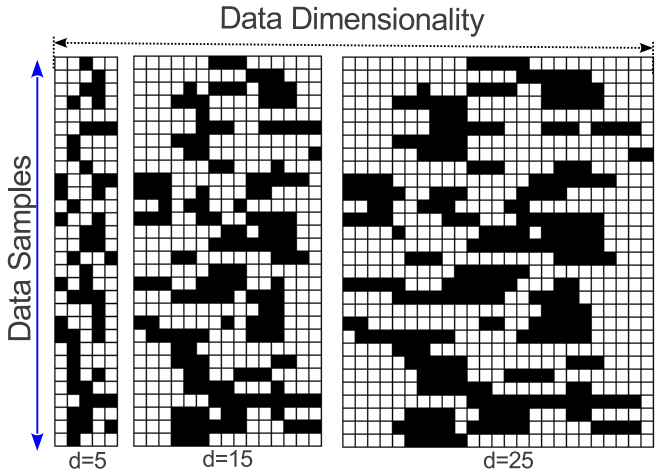

Fig. 4. First thirty samples of three different resolutions of the artificial 0-1 data. Black denotes 1s and white denotes 0s.

First, we create a matrix of the required size with all zeros. We divide the unit interval in 25 equal parts to generate 25–dimensional data. The parameter, X, defines the beginning of an aberration, and l defines the length of the aberration. We randomly choose a data sample for aberration and flip zeros to ones from dimensionality X of length l i.e. from dimension X to dimension $X+l$. We ignore the the lengths that are greater than the dimension $25 < (X+l)$ to maintain the data dimensionality. We continue this process iteratively until the number of 1s is approximately 55-60% of the data to mimic chromosomal aberration datasets.

Domain knowledge of chromosomal aberrations informs us about the typical length of aberrations. Furthermore, aberrations never span across the centromere. Therefore, we break an aberration that is longer than a predefined length, 15 in the experiments, randomly either on the left or the right of the centromere. We fix the centromere after the 10^{th} dimension, i.e. variable on the 10^{th} dimension is on the left side of the centromere, and the variable on the 11^{th} dimension is on the right side of the centromere.

Figure 4 shows artificial data in three different resolutions with dimensionalities 5, 15, and 25, respectively. Figure 4 also shows that similar to chromosomal aberration data, the artificial data are sparse, and spatially dependent. We gain the knowledge of relationships between data in different resolutions from the

Bayesian network as shown in the middle panel (within the dotted rectangle) of Figure 3. We apply that knowledge to downsample the data in dimensionality 25 to a dimensionality of 15, and then 5 using the majority voting downsampling method proposed in [14]. In experiments we fix the number of samples of the dataset to 1000 which is similar to the chromosomal aberration dataset.

4.2 Multiresolution Chromosomal Aberration Dataset

The causes and consequences of chromosomal aberration such as amplification, deletion, and duplication have significant roles in cancer research [13]. DNA copy number amplifications have been defined as the hallmarks of cancer [11,28]. Chromosomal aberrations are early markers of cancer risk and their detection and analysis has considerable clinical merits [11,29]. The two chromosomal aberration datasets in two resolutions have a dimensionality of 393, and 862 in coarse, and fine resolution, respectively. Both datasets are used in [11,27,28]. Datasets are available from the authors on request. The sources of the two datasets were different and correspond to different cancer patients in the two different resolutions.

We experiment chromosome–wise to constrain the complexity of learning the mixture model because the data are high dimensional and samples are small. Complexity of mixture models increases quadratically with the dimensionality. For example, the number of samples in chromosome 17 is 342 and 2716 in coarse and fine resolution, respectively. Therefore, we correspond the first 342 samples in fine resolution to the samples in coarse resolution. We then downsample the next 342 samples (samples 343 to 684) to a resolution between coarse and fine resolution such that the depth and structure of the resulting Bayesian network are similar to those of the networks used for artificial dataset. We ignore the remaining 2032 samples in the fine resolution.Similarly, the network present in the real world dataset differ from the artificial dataset. We select the most representative network covering more than 50% of the data and ignore the other networks. The structures of the networks are similar (often number of types of trees in the dataset is about ≈ 3).

5 Experiments and Results

The experimental studies in this paper are a two-step procedure because the algorithm models multiresolution data in two steps. Firstly, we learn the component distributions in a Bayesian network framework from different resolutions of the data. Secondly, we model multiresolution data after transforming the Bayesian networks to vectors using mixture models.

5.1 Experiments with Bayesian Networks of Multiresolution Data

From the knowledge of multiresolution data relationships in Section 4, we generate the five different Bayesian networks abreast adjacently along the dimensionality of the data. We use BRMLToolbox to encode and generate Bayesian

network [22]. We use a maximum likelihood framework to learn the conditional probabilities of the network.

As discussed in Section 2.2, some of the resolutions (entire datasets) can be missing in a multiresolution scenario. We use the prowess of Bayesian networks in handling missing data [12] to impute missing resolutions. In experimental setup: firstly, we learn the parameters of the component distributions in Bayesian network framework via maximum likelihood. We then ascertain the performance of component distributions as Bayesian network models especially with respect to their ability to impute missing values. We artificially generate two scenarios: where one, and two resolutions of data are missing. We draw the samples from a consistent junction tree in the Bayesian network under the given evidence using BRMLToolbox [22]. The number of samples equal that of the original dataset. We then compare the re-sampled data with the original data.

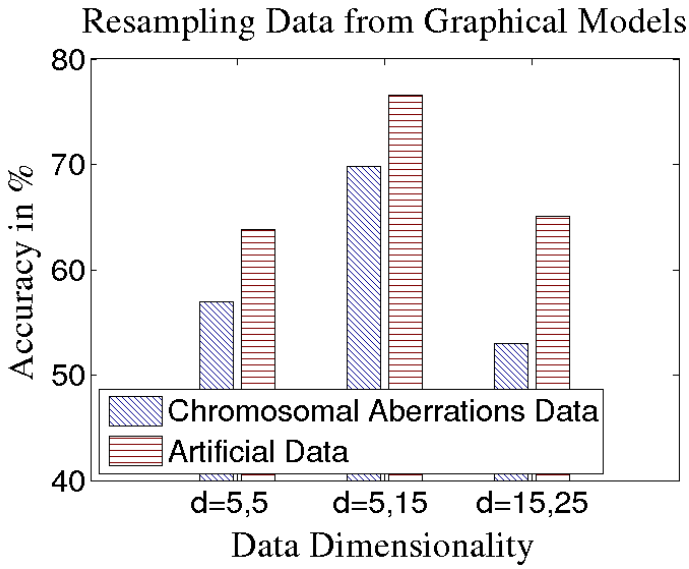

Fig. 5. The accuracy in re-sampling the data in missing resolutions conditioned on the data in other available resolutions. Comma in the X-axis separates the dimensionalities of the artificial data and the chromosomal aberration data, i.e. d=x,y denotes dimensionality (d) = (chromosomal aberrations data dimension (x), artificial data dimension (y)).

We calculate the matrix difference between the original data and the data sampled from the Bayesian networks. The difference in the binary data is sum of the number of places where 0s are 1s, and 1s are 0s. The difference is comparatively small in smaller dimensions than in the larger dimensions because the cumulative difference depends on the size of the dataset. Therefore, we calculated the accuracy of element-wise matching of two datasets as shown in Figure 5. Accuracy is the percentage of places where each element of both the matrices are equal.

In artificial data, when the data resolution with dimensionality 15 is missing, other data resolutions with dimensionality 5, and 25 are available, and we need to impute only data in resolution 15. Similarly, Data resolutions with dimensionalities 5, and 25 are missing when the data resolution with dimensionality 15 is available. Therefore, we should impute both the resolutions with dimensionalities 5, and 15. In this case, we can simply run single resolution analysis. However, we try to artificially create this scenario to demonstrate that our algorithm performs reliably under harsh conditions of large amount of missing values. In chromosomal aberrations data, coarse and the in–between data resolutions have the same dimensionality of 5. Therefore, when the in–between data resolution with dimensionality 5 is missing, coarse, and fine data resolution with dimensionality 5, and 15 are available, respectively. Similarly, the coarse, and fine data resolutions with dimensionalities 5, and 15 are missing when the data in the in–between resolution with dimensionality 5 are available.

The results in Figure 5 show that accuracy of matching is higher when two resolutions of data are available and only the data in a single resolution are missing. When only one data resolution is available, and we need to impute two resolutions in the coarse and the fine resolution, the accuracy is poorer. This result is intuitive because the number of known variables is smaller than the number of missing variables when two data resolutions are missing. Similarly, accuracy is poorer in high dimensional data (fine resolution) compared to data with lower dimensionality (coarse resolution). This discrepancy is the result of the curse of dimensionality phenomenon. Overall, the results show that the model of component distributions as Bayesian networks produces plausible results.

5.2 Experiments on Mixture Modeling of Multiresolution Data

In experimental setup, firstly, we transform the multiresolution data to the Bayesian network representation as shown in the Figure 3. Secondly, we transform the Bayesian network representation to the vector representation after imputing missing values (if any) as explained in Section 3. In the vector representation, the transformed multiresolution data have same dimensionality. The EM algorithm learns the mixture model with a priori knowledge of the number of components for data. As in [11,27,28], we use model selection in a 10-fold cross-validation setting to select the appropriate number of mixture components.

We train models of different complexities in a ten-fold cross–validation setting and select the model with the best generalization performance. Figure 6 shows that both training and validation likelihood steadily increases until the number of components is 5, then smoothen and flatten after the number of components is 5. This suggests that 5 is the appropriate number of mixture components.

After selecting the number of components, we train 200 different models of the same complexity and choose the model that produces the best likelihood on the data to ameliorate the problem of local optima in the EM algorithm. We also perform similar experiments with data in each resolution to select the number of mixture components and train the mixture model as a comparison with the results of the multiresolution model. Table 1 shows the variation in the number

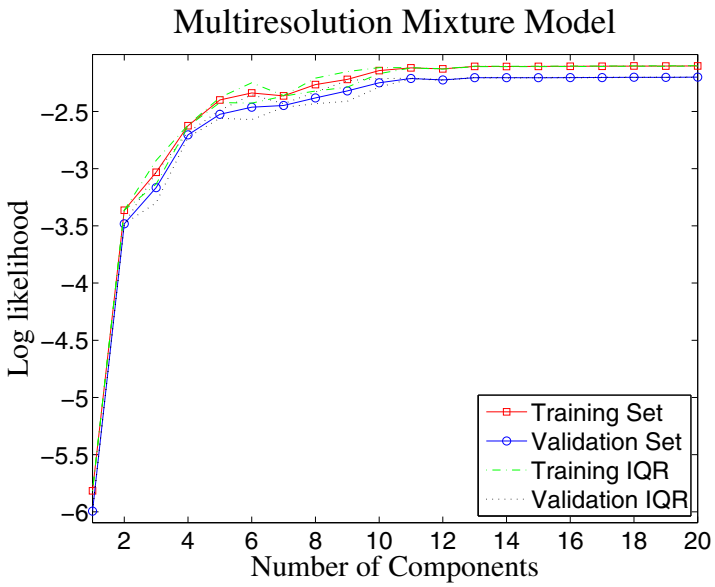

Fig. 6. Model selection in a 10-fold cross-validation setting in multiresolution artificial data. Final results for the same model selection are denoted by a boldfaced row in Table 1. Averaged training and validation likelihood along with their corresponding Inter Quartile Range (IQR) for each training and validation run has also been plotted. The selected number of components is 5.

Table 1. The results of mixture modeling on single resolution and multiresolution models. Here, J and $\mathcal{L}$, denote the selected number of components and the log likelihood obtained by the best model, respectively.

Artificial Data			**Chromosome-17**		
Datasets (Dimension)	**Results**		**Datasets Dimension)**	**Results**	
	J	$\mathcal{L}$		**J**	$\mathcal{L}$
Single Resolution (5)	3	-3.24	Single Resolution (5)	4	-2.23
Single Resolution (15)	6	-8.32	Single Resolution (5)	3	-2.17
Single Resolution (25)	7	-12.84	Single Resolution (15)	5	-3.73
Multiresolution (9)	**5**	**-2.40**	Multiresolution (5)	4	-2.14

of components required to fit the data in different resolutions. Furthermore, the likelihood is considerably smaller in single resolution showing improvement in mixture modeling because of the use of multiple resolutions.

Figure 7 shows the log likelihood of three single resolution models and a multiresolution model. We trained the mixture model initialized at random in a ten-fold cross-validation setting with the selected number of components to convergence, i.e. until the increase in log-likelihood is small, 0.0001 in the experiments. The shorter the bar the better the result as Y-axis depicts negative log likelihood.

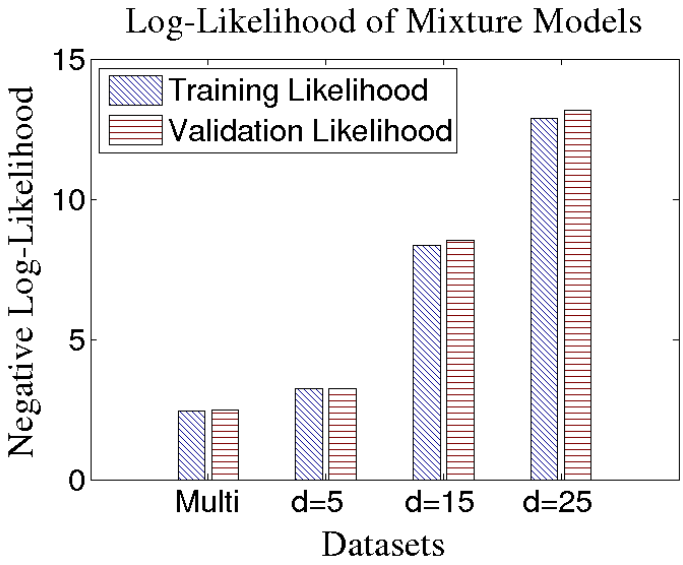

Fig. 7. The log likelihood of three mixture models in single resolution and a multiresolution mixture model trained in a 10-fold cross-validation setting after selecting the number of components. The Y-axis shows the negative log likelihood, therefore, the shorter the bar, the better the result.

We select a different number of components for each dataset as shown in Table 1. The results also show that multiresolution mixture model outperforms the single resolution models. Log likelihood is comparatively smaller in dimensionalities of 15, and 25 because of the increased dimensionality of the data. The likelihood of the multiresolution model is better than the data with the smallest dimensionality of 5 in single resolution although the dimensionality of the multiresolution data is 9. The Table 1 also shows similar results on chromosomal data.

6 Summary and Conclusions

In this paper, we proposed a mixture model of multiresolution components to model multiresolution 0-1 data. Firstly, we design the multiresolution components of the mixture model as Bayesian networks with the knowledge of the hierarchy of resolutions from the domain ontology. We then learn the CPD of the networks from the multiresolution data. Secondly, we transform the multiresolution component distributions to vector representation and learn the mixture model in a ten-fold cross validation setting. We experimented with the algorithm on a multiresolution artificial dataset and also on a multiresolution chromosomal aberration dataset. The experimental results show that the proposed approach of multiresolution modeling outperforms single resolution models.

Acknowledgments. The work is funded by Helsinki Doctoral Programme in Computer Science – Advanced Computing and Intelligent Systems (**Hecse**), and Finnish Centre of Excellence for Algorithmic Data Analysis Research (**ALGODAN**). I would also like to thank colleague Mikko Korpela for reading through the manuscript and suggesting the improvements in presentation.

References

1. Garland, M.: Multiresolution Modeling: Survey & Future Opportunities. In: Eurographics 1999 – State of the Art Reports, pp. 111–131 (1999)
2. Willsky, A.S.: Multiresolution Markov Models for Signal and Image Processing. Proceedings of the IEEE 90(8), 1396–1458 (2002)
3. Shaffer, L.G., Tommerup, N.: ISCN 2005: An International System for Human Cytogenetic Nomenclature(2005) Recommendations of the International Standing Committee on Human Cytogenetic Nomenclature. Karger (2005)
4. Lindeberg, T.: Scale-space theory: A basic tool for analysing structures at different scales. Journal of Applied Statistics 21(2), 224–270 (1994)
5. Vetterli, M., Kovačevic, J.: Wavelets and Subband Coding. Prentice-Hall, Inc., Upper Saddle River (1995)
6. Russell, B.: On the Relations of Universals and Particulars. Proceedings of the Aristotelian Society 12, 1–24 (1911)
7. Everitt, B.S., Hand, D.J.: Finite Mixture Distributions. Chapman and Hall, London (1981)
8. McLachlan, G.J., Peel, D.: Finite Mixture Models. Probability and Statistics – Applied Probability and Statistics Section, vol. 299. Wiley, New York (2000)
9. Moore, A.: Very Fast EM-based Mixture Model Clustering Using Multiresolution KD–trees. In: Kearns, M., Cohn, D. (eds.) Advances in Neural Information Processing Systems, pp. 543–549. Morgan Kaufmann (April 1999)
10. Meilâ, M., Jordan, M.I.: Learning with mixtures of trees. Journal of Machine Learning Research 1, 1–48 (2000)
11. Myllykangas, S., Tikka, J., Böhling, T., Knuutila, S., Hollmén, J.: Classification of human cancers based on DNA copy number amplification modeling. BMC Medical Genomics 1(15) (May 2008)
12. Marlin, B.M.: Missing data problems in machine learning. PhD thesis, University of Toronto (2008)
13. Kirsch, I.R.: The Causes and Consequences of Chromosomal Aberrations, 1st edn. CRC Press (December 1992)
14. Adhikari, P.R., Hollmén, J.: Patterns from multiresolution 0-1 data. In: Proceedings of the ACM SIGKDD Workshop on Useful Patterns, UP 2010, pp. 8–16. ACM, New York (2010)
15. Adhikari, P.R., Hollmén, J.: Multiresolution Mixture Modeling using Merging of Mixture Components. In: Hoi, S.C.H., Buntine, W. (eds.) Proceedings of the Fourth Asian Conference on Machine Learning, ACML 2012, JMLR Workshop and Conference Proceedings, Singapore, vol. 25, pp. 17–32 (2012)
16. Wilson, R.: MGMM: multiresolution Gaussian mixture models for computer vision. In: Proceedings of 15th International Conference on Pattern Recognition, vol. 1, pp. 212–215 (2000)

17. Ng, S.-K., McLachlan, G.J.: Robust Estimation in Gaussian Mixtures Using Multiresolution Kd-trees. In: Sun, C., Talbot, H., Ourselin, S., Adriaansen, T. (eds.) Proceedings of the 7th International Conference on Digital Image Computing: Techniques and Applications, pp. 145–154. CSIRO Publishing (2003)
18. Bellot, D.: Approximate discrete probability distribution representation using a multi–resolution binary tree. In: Proceedings of 15th IEEE International Conference on Tools with Artificial Intelligence, pp. 498–503 (2003)
19. Sanchís, F.A., Aznar, F., Sempere, M., Pujol, M., Rizo, R.: Learning Discrete Probability Distributions with a Multi-resolution Binary Tree. In: Corchado, E., Yin, H., Botti, V., Fyfe, C. (eds.) IDEAL 2006. LNCS, vol. 4224, pp. 472–479. Springer, Heidelberg (2006)
20. Bianchini, M., Maggini, M., Sarti, L.: Object Recognition Using Multiresolution Trees. In: Yeung, D.-Y., Kwok, J.T., Fred, A., Roli, F., de Ridder, D. (eds.) SSPR&SPR 2006. LNCS, vol. 4109, pp. 331–339. Springer, Heidelberg (2006)
21. Huerta, J., Chover, M., Quiros, R., Vivo, R., Ribelles, J.: Binary space partitioning trees: a multiresolution approach. In: Proceedings of 1997 IEEE Conference on Information Visualization, pp. 148–154 (1997)
22. Barber, D.: Bayesian Reasoning and Machine Learning. Cambridge University Press (2012)
23. Jordan, M.I.: Graphical Models. Statistical Science (2004)
24. Heckerman, D.: A Tutorial on Learning With Bayesian Networks. In: Jordan, M.I. (ed.) Learning in Graphical Models, pp. 301–354. MIT Press, USA (1999)
25. Enders, C.K.: Applied Missing Data Analysis, 1st edn. The Guilford Press (2010)
26. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39(1), 1–38 (1977)
27. Adhikari, P.R., Hollmén, J.: Fast Progressive Training of Mixture Models for Model Selection. In: Ganascia, J.-G., Lenca, P., Petit, J.-M. (eds.) DS 2012. LNCS (LNAI), vol. 7569, pp. 194–208. Springer, Heidelberg (2012)
28. Tikka, J., Hollmén, J., Myllykangas, S.: Mixture Modeling of DNA copy number amplification patterns in cancer. In: Sandoval, F., Prieto, A.G., Cabestany, J., Graña, M. (eds.) IWANN 2007. LNCS, vol. 4507, pp. 972–979. Springer, Heidelberg (2007)
29. Lu, X., Shaw, C.A., Patel, A., Li, J., Cooper, M.L., Wells, W.R., Sullivan, C.M., Sahoo, T., Yatsenko, S.A., Bacino, C.A., Stankiewicz, P., Ou, Z., Chinault, A.C., Beaudet, A.L., Lupski, J.R., Cheung, S.W., Ward, P.A.: Clinical Implementation of Chromosomal Microarray Analysis: Summary of 2513 Postnatal Cases. PLoS ONE 2(3), e327 (2007)

Publication V

Prem Raj Adhikari, Anže Vavpetič, Jan Kralj, Nada Lavrač, Jaakko Hollmén. Explaining mixture models through semantic pattern mining and banded matrix visualization. In *Proceedings of Seventeenth International Conference on Discovery Science (DS 2014)*, Sašo Džeroski, Panče Panov, Dragi Kocev, Ljupčo Todorovski, Editors, Volume 8777 of Lecture Notes in Computer Science, Springer International Publishing Switzerland 2014, pages 1-12, October 8–10, 2014, Bled, Slovenia. DOI: 10.1007/978-3-319-11812-3_1, October 2014.

Explaining Mixture Models through Semantic Pattern Mining and Banded Matrix Visualization

Prem Raj Adhikari[1], Anže Vavpetič[2], Jan Kralj[2], Nada Lavrač[2], and Jaakko Hollmén[1]

[1] Helsinki Institute for Information Technology HIIT and Department of Information and Computer Science, Aalto University School of Science, PO Box 15400, FI-00076 Aalto, Espoo, Finland
{prem.adhikari,jaakko.hollmen}@aalto.fi

[2] Jožef Stefan Institute and Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
{anze.vavpetic,jan.kralj,nada.lavrac}@ijs.si

Abstract. Semi-automated data analysis is possible for the end user if data analysis processes are supported by easily accessible tools and methodologies for pattern/model construction, explanation, and exploration. The proposed three–part methodology for multiresolution 0–1 data analysis consists of data clustering with mixture models, extraction of rules from clusters, as well as data, cluster, and rule visualization using banded matrices. The results of the three-part process—clusters, rules from clusters, and banded structure of the data matrix—are finally merged in a unified visual banded matrix display. The incorporation of multiresolution data is enabled by the supporting ontology, describing the relationships between the different resolutions, which is used as background knowledge in the semantic pattern mining process of descriptive rule induction. The presented experimental use case highlights the usefulness of the proposed methodology for analyzing complex DNA copy number amplification data, studied in previous research, for which we provide new insights in terms of induced semantic patterns and cluster/pattern visualization.

Keywords: Mixture Models, Semantic Pattern Mining, Pattern Visualization.

1 Introduction

In data analysis, the analyst aims at finding novel ways to summarize the data to become easily understandable [6]. The interpretation aspect is especially valued among application specialists who may not understand the data analysis process itself. Semi-automated data analysis is hence possible for the end user if data analysis processes are supported by easily accessible tools and methodologies for pattern/model construction, explanation and exploration. This work draws together

S. Džeroski et al. (Eds.): DS 2014, LNAI 8777, pp. 1–12, 2014.

different approaches developed in our previous research, leading to a new three-part data analysis methodology. Its utility is illustrated in a case study concerning the analysis of DNA copy number amplifications represented as a 0–1 (binary) dataset [12]. In our previous work we have successfully clustered this data using mixture models [13, 16]. Furthermore, in [8], we have learned linguistic names for the patterns that coincide with the natural structure in the data, enabling domain experts to refer to clusters or the patterns extracted from these clusters, with their names. In [7] we report that frequent itemsets describing the clusters, or extracted from the 'one cluster at a time' clustered data are markedly different than those extracted from the whole dataset. The whole set of about 100 DNA amplification patterns identified from the data have been reported in [13].

With the aim of better explaining the initial mixture model based clusters, in this work we consider the cluster labels as class labels in descriptive rule learning [14], using a semantic pattern mining approach [20]. This work proposes a crossover of unsupervised methods of probabilistic clustering with supervised methods of subgroup discovery to determine the specific chromosomal locations, which are responsible for specific types of cancers. Determining the chromosomal locations and their relation to certain cancers is important to study and understand pathogenesis of cancer. It also provides necessary information to select the optimal target for cancer therapy on individual level [10]. We also enrich the data with additional background knowledge in different forms such as pre–discovered patterns as well as taxonomies of features in multiresolution data, cancer genes, and chromosome fragile sites. The background knowledge enables the analysis of data at multiple resolutions. This work reports the results of the Hedwig semantic pattern mining algorithm [19] performing semantic subgroup discovery, using the incorporated background knowledge.

While a methodology, consisting of clustering and semantic pattern mining, has been suggested in our previous work [9, 20], we have now for the first time addressed the task of explaining sub-symbolic mixture model patterns (clusters of instances) using symbolic rules. In this work, we propose this two-step approach to be enhanced through pattern comparison by their visualization on the plots resulting from banded matrices visualization [5]. Using colored overlays on the banded patient–chromosome matrix (induced from the original data), the mixture model clusters are first visualized, followed by visualizing the sets of patterns (i.e., subgroups) induced by semantic pattern mining.

Matrix visualization is a very popular method for information mining [1] and banded matrix visualization provides new means for data and pattern exploration, visualization and comparison. The addition of visualization helps to determine if the clustering results are plausible or awry. It also helps to identify the similarities and differences between clusters with respect to the amplification patterns. Moreover, an important contribution of this work is the data analytics task addressed, i.e., the problem of explaining chromosomal amplification in cancer patients of 73 different cancer types where data features are represented in multiple resolutions. Data is generated in multiple resolutions (different dimensionality) if a phenomenon is measured in different levels of detail.

The main contributions of this work are as follows. We propose a three-part methodology for data analysis. First, we cluster the data. Second, we extract semantic patterns (rules) from the clusters, using an ontology of relationships between the different resolutions of the multiresolution data [15]. Finally, we integrate the results in a visual display, illustrating the clusters and the identified rules by visualizing them over the banded matrix structure.

2 Methodology

A pipeline of algorithmic steps forming the proposed three–part methodology is outlined in Figure 1. The methodology starts with a set of experimental data (*Load Data*) and background knowledge and facts (*Load Background Knowledge*) as shown in the Figure 1. Next, both a mixture model (*Compute Mixture Model*) and a banded matrix (*Compute Banded Matrix*) are induced independently from the data in Sections 3.1 and 3.2, respectively. The mixture model is then applied to the original data, to obtain a clustering of the data (*Apply Mixture Model*). The banded structure enables the visualization of the resulting clusters in Section 3.2 (*Banded matrix cluster visualization*). Semantic pattern mining is used in Section 3.3 to describe the clusters in terms of the background knowledge (*Semantic pattern mining*). Finally, all three models (the mixture model, the banded matrix and the patterns) are joined in Section 3.4 to produce the final visualization (*Banded Matrix Rule Visualization*).

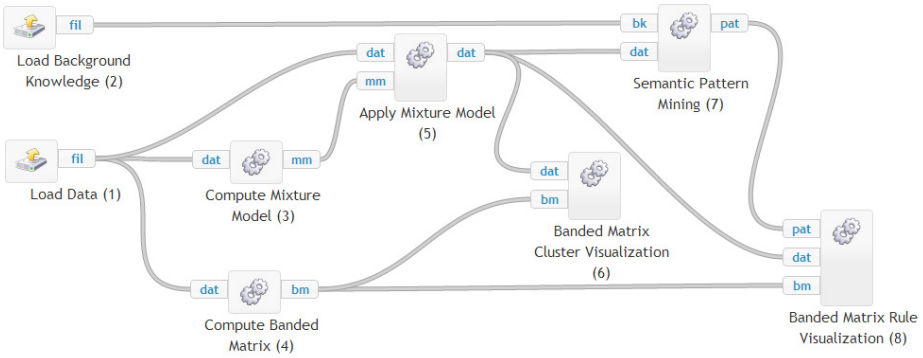

Fig. 1. Overview of the proposed three-part methodology

2.1 Experimental Data

The dataset under study describes DNA copy number amplifications in 4,590 cancer patients. The data describes 4,590 patients as data instances, with attributes being chromosomal locations indicating aberrations in the genome. These aberrations are described as 1's (amplification) and 0's (no amplification). Authors in [12] describe the amplification dataset in detail. In this paper, we consider datasets in different resolutions of chromosomal regions as defined by International System of Cytogenetic Nomenclature (ISCN) [15].

Given the complexity of the multiresolution data, we were forced to reduce the complexity of the learning setting to a simpler setting, allowing us to develop and test the proposed methodology. To this end, we have reduced the size of the dataset: from the initial set of instances describing 4,590 patients, each belonging to one of the 73 different cancer types, we have focused on 34 most frequent cancer types only, as there were small numbers of instances available for many of the rare cancer types, thus reducing the dataset from 4,590 instances to a 4,104 instances dataset. In addition, in the experiments we have focused on a single chromosome (chromosome 1), using as input to step 2 of the proposed methodology the data clusters obtained at the 393 locations granularity level using a mixture modeling approach [13]. Inferencing and density estimation from entire data would produce degenerate results because of the curse of large dimensionality. When chromosome 1 is extracted from the data, some cancer patients show no amplifications in any bands of the chromosome 1. We have removed such samples without amplifications (zero vectors) because we are interested in the amplifications and their relation to cancers, not their absence. This reduces the sample size of chromosome 1 from 4,104 to 407. Similar experiments can be performed for each chromosome in such a way that every sample of data is properly utilized. While this data reduction may be an over-simplification, finding relevant patterns in this dataset is a huge challenge, given the fact that even individual cancer types are known to consist of cancer sub-types which have not yet been explained in the medical literature. The proposed methodology may prove, in future work, to become a cornerstone in developing means through which such sub-types could be discovered, using automated pattern construction and innovative pattern visualization using banded matrices visualization.

In addition to the DNA amplifications dataset, we used supplementary background knowledge in the form of an ontology to enhance the analysis of the dataset. The supplementary background knowledge used are taxonomies of hierarchical structure of multiresolution amplification data, chromosomal locations of fragile sites [3], virus integration sites [21], cancer genes [4], and amplification hotspots. The hierarchical structure of multiresolution data is due to ISCN which allows the exact description of all numeric and structural amplifications in genomes [15]. Amplification hotspots are frequently amplified chromosomal loci identified using computational modeling [12].

2.2 Mixture Model Clustering

Mixture models are probabilistic models for modeling complex distributions by a mixture or weighted sum of simple distributions through a decomposition of the probability density function into a set of component distributions [11]. Since the dataset of our interest is a 0–1 data, we use multivariate Bernoulli distributions as component distributions to model the data. Mathematically, this can be expressed as

$$p(\boldsymbol{x}|\boldsymbol{\Theta}) = \sum_{j=1}^{\mathrm{J}} \pi_j P(\mathbf{x} \mid \theta_j) = \sum_{j=1}^{\mathrm{J}} \pi_j \prod_{i=1}^{\mathrm{d}} \theta_{ji}^{x_i} (1-\theta_{ji})^{1-x_i}. \quad (1)$$

Here, $j = 1, 2, \ldots, J$ indexes the component distributions and $i = 1, 2, \ldots, d$ indexes the dimensionality of the data. π_j defines the mixing proportions or mixing coefficients determining the weight for each of the J component distributions. Mixing proportions satisfy the properties of convex combination such as: $\pi_j \geq 0$ and $\sum_{J=1}^{J} \pi_j = 1$. Individual parameters θ_{ji} determine the probability that a random variable in the j^{th} component in the i^{th} dimension takes the value 1. x_i denotes the data point such that $x_i \in \{0, 1\}$. Therefore, the parameters of mixture models can be represented as: $\boldsymbol{\Theta} = \{J, \{\pi_j, \Theta_j\}_{j=1}^{J}\}$.

Expectation maximization (EM) algorithm can be used to learn the maximum likelihood parameters of the mixture model if the number of component distributions are known in advance [2]. Whereas the mixture model is merely a way to represent the probability distribution of the data, the model can be used in clustering the data into (hard) partitions, or subsets of data instances. We can achieve this by allocating individual data vectors to mixture model components that maximize the posterior probability of that data vector.

2.3 Semantic Pattern Mining

Existing semantic subgroup discovery algorithms are either specialized for a specific domain [17] or adapted from systems that do not take into the account the hierarchical structure of background knowledge [18]. The general purpose Hedwig system overcomes these limitations by using domain ontologies to structure the search space and formulate generalized hypotheses. Semantic subgroup discovery, as addressed by the Hedwig system, results in relational descriptive rules. Hedwig uses ontologies as background knowledge and training examples in the form of Resource Description Framework (RDF) triples. Formally, we define the semantic data mining task addressed in the current contribution as follows.

Given:
- set of training examples in empirical data expressed as RDF triples
- domain knowledge in the form of ontologies, and
- an object-to-ontology mapping which associates each object from the RDF triplets with appropriate ontological concepts.

Find:
- a hypothesis (a predictive model or a set of descriptive patterns), expressed by domain ontology terms, explaining the given empirical data.

The Hedwig system automatically parses the RDF triples (a graph) for the `subClassOf` hierarchy, as well as any other user-defined binary relations. Hedwig also defines a namespace of classes and relations for specifying the training examples to which the input must adhere. The rules generated by Hedwig system using beam search are repetitively specialized and induced as discussed in [20].

The significance of the findings is tested using the Fisher's exact test. To cope with the multiple–hypothesis testing problem, we use Holm–Bonferroni direct adjustment method with $\alpha = 0.05$.

2.4 Visualization Using Banded Matrices

Consider a $n \times m$ binary matrix $M$and two permutations, κ and π of the first n and m integers. Matrix M_κ^π, defined as $(M_\kappa^\pi)_{i,j} = M_{\kappa(i),\pi(j)}$, is constructed by applying the permutations π and κ on the rows and columns of M. If, for some pair of permutations π and κ, matrix M_κ^π has the following property:

- For each row i of the matrix, the column indices for which the value in the matrix is 1 appear consecutively, i.e., on indices $a_i, a_i + 1, \ldots, b_i$,
- For each i, we have $a_i \leq a_{i+1}$ and $b_i \leq b_{i+1}$,

then the matrix M is *fully banded* [5]. The motivation behind using banded matrices to exposes the clustered structure of the underlying data through the banded structure. We use barycentric method used to extract banded matrix in [5] to find the banded structure of a matrix. The core idea of the method is the calculation of barycenters for each matrix row, which are defined as

$$Barycenter(i) = \frac{\sum_{j=1}^{m} j \cdot M_{ij}}{\sum_{j=1}^{m} M_{ij}}.$$

The barycenters of each matrix row are best understood as centers of gravity of a stick divided into m sections corresponding to the row entries. An entry of 1 denotes a weight on that section. One step of the `barycentric` method now: calculate the barycenters for each matrix row and sort the matrix rows in order of increasing barycenters. In this way, the method calculates the best possible permutation of rows that exposes the banded structure of the input matrix. It does not, however, find any permutation of columns. In our application, neighboring columns of a matrix represent chromosome regions that are in physical proximity to one another, the goal is to only find the optimal row permutation while not permuting the matrix columns.

The image of the banded structure can then be overlaid with a visualization of clusters, as described in Section 2.2. Because the rows of the matrix represent instances, highlighting one set of instances (one cluster) means highlighting several matrix rows. If the discovered clusters are exposed by the matrix structure, we can expect that several adjacent matrix rows will be highlighted, forming a wide band. Furthermore, all the clusters can be simultaneously highlighted because each sample belongs to one and only one cluster. The horizontal colored overlay of the clusters in Figure 3 can also be supplemented with another colored vertical overlay of the rules explaining the clusters as discussed in Section 2.3. If an important chromosome region is discovered for the characterization of a cluster, we highlight the corresponding column. In the case of composite rules of the type Rule 1: `Cluster3(X)` $\leftarrow$ `1q43-44(X)` $\wedge$ `1q12(X)`, both chromosomal regions $1q43--44$ and $1q12$ are understood as equally important and are

therefore both highlighted. If a chromosome band appears in more than one rule, this is visualized by a stronger highlight of the corresponding matrix column.

3 Experiments and Results

3.1 Clusters from Mixture Models

We used the mixture models trained in our earlier contribution [13]. Through a model selection procedure documented in [16], the number of components for modeling the chromosome 1 had been chosen to be $J = 6$, denoting presence of six clusters in the data.

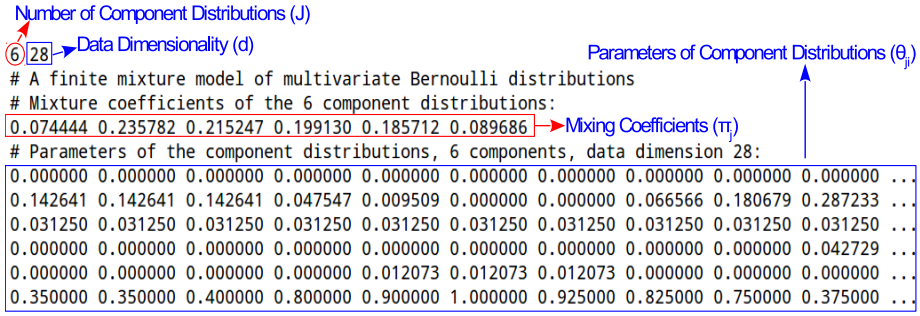

Fig. 2. Mixture model for chromosome 1. Figure shows first 10 dimensions of the total 28 for clarity.

Figure 2 shows a visual illustration of the mixture model for chromosome 1. In the figure, the first line denotes the number of components (J) in the mixture model and the data dimensionality (d). The lines beginning with # are comments and can be ignored. The fourth line shows the parameters of component distributions (π_j) which are six probability values summing to 1. Similarly, the last six lines of the figure denote the parameters of the component distributions, θ_{ji}. Figure 2 does not visualize and summarize the data as it consists of many numbers and probability values. Therefore, we use banded matrix for visualization as discussed in Section 2.4. Here, we focus on hard clustering of the samples of chromosomal amplification data using the mixture model depicted in Figure 2. The dataset is partitioned into six different clusters allocating data vectors to the component densities that maximize the probability of data. The number of samples in each cluster are the following: Cluster 1→30, Cluster 2→96, Cluster 3→88, Cluster 4→81, Cluster 5→75, and Cluster 6→37.

3.2 Cluster Visualization Using Banded Matrices

We used the barycentric method, described in Section 2.4, to extract the banded structure in the data. The black color indicates ones and white color denotes zeros in the data. The banded data was then overlaid with different colors for

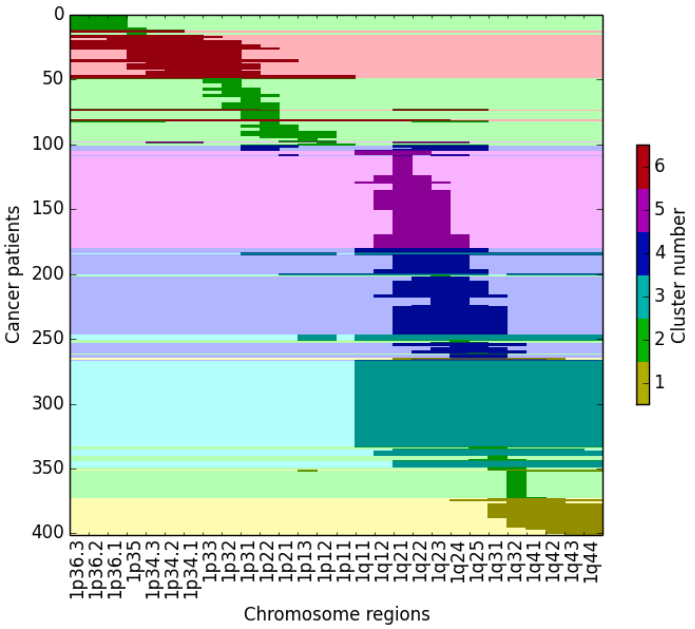

Fig. 3. Banded structure of the matrix with cluster information overlay

the 6 clusters, discovered in Section 3.1, as shown in the Figure 3. By exposing the banded structure of the matrix, Figure 3 allows a clear visualization of the clusters discovered in the data. Figures 3 shows that each cluster captures amplifications in some specific regions of the genome. The figure captures a phenomenon that the left part of the figure showing chromosomal regions beginning with $1p$ shows a comparatively smaller number of amplifications whereas the right part of the figure showing chromosomal regions beginning with $1q$ (q–arm) shows a higher number of amplifications.

The Figure 3 also shows that cluster 1 is characterized by pronounced amplifications in the end of the q–arm (regions 1q32–q44) of chromosome 1. The figure also shows that samples in the second cluster contain sporadic amplifications spread across both p and q–arms in different regions of chromosome 1. This cluster does not carry much information and contains cancer samples that do not show discriminating amplifications in chromosomes as the values of random variables are near 0.5. It is the only cluster that was split into many separate matrix regions. In contrast, cluster 3 portrays marked amplifications in regions 1q11–44. Cluster 4 shows amplifications in regions 1q21–25. Similarly, cluster 5 is denoted by amplifications in 1q21–25. Cluster 6 is defined by pronounced amplifications in the p-arm of chromosome 1. The visualization with banded matrices in Figure 3 also draws a distinction between clusters each cluster which upon first viewing show no obvious difference to the human eye when looking at the probabilities of the mixture model shown in the Figure 2.

3.3 Rules Induced Through Semantic Pattern Mining

Using the method described in Section 2.3, we induced subgroup descriptions for each cluster as the target class [19]. For a selected cluster, all the other clusters represent the negative training examples, which resembles a one-versus-all approach in multiclass classification. In our experiments, we consider only the rules without negations, as we are interested in the presence of amplifications characterizing the clusters (and thereby the specific cancers), while the absence of amplifications normally characterizes the absence of cancers not their presence [10]. We focus our discussion only on the results pertaining to cluster 3 because of the space constraints.

Rules for cluster 3	Precision	Lift
C3(X) ← 1q43-44(X) ∧ 1q12(X)	1.00	4.62
C3(X) ← 1q11(X)	0.90	4.15
C3(X) ← 1q43-44(X)	0.77	3.57
C3(X) ← 1q41(X)	0.76	3.51
C3(X) ← 1q12(X)	0.65	3.02
C3(X) ← 1q32(X)	0.63	2.91
C3(X) ← 1q31(X)	0.62	2.85
C3(X) ← 1q25(X)	0.58	2.68
C3(X) ← 1q24(X)	0.48	2.20
C3(X) ← 1q21(X)	0.40	1.83
C3(X) ← 1q22--24(X)	0.37	1.72
C3(X) ← HotspotSite(X)	0.28	1.31
C3(X) ← CancerSite(X)	0.26	1.22
C3(X) ← FragileSite(X)	0.25	1.17

Table 1: Rules induced for clusters 3.

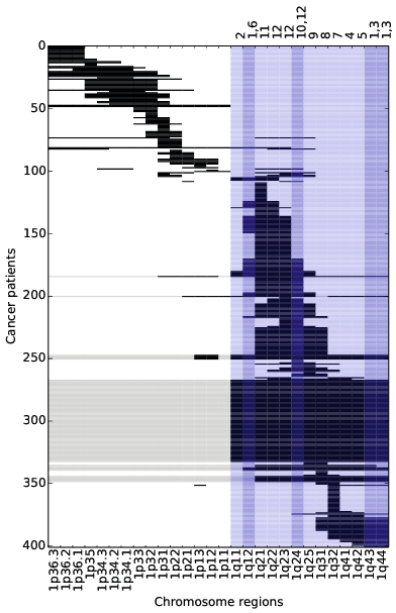

Fig. 4. Rules induced for cluster 3 (left) and visualizations of rules and columns for cluster 3 (right) with relevant columns highlighted. A highlighted column denotes that an amplification in the corresponding region characterizes the instances of the particular cluster. A darker hue means that the region appears in more rules. The numbers on top right correspond to rule numbers. For example, the notations "1, 3" on top of rightmost column of cluster 3 indicates that the chromosome region appears in rules 1 and 3 tabulated in the left panel.

Table on the left panel of the Figure 4 show the rules induced for cluster 3, together with the relevant statistics. The induced rules quantify the clustering results obtained in Section 3.1 and confirmed by banded matrix visualization in Section 3.2. The banded matrix visualization depicted in Figure 3 shows that cluster 3 is marked by the amplifications in the regions 1q11–44. However, the rules obtained in Table on the left panel of the Figure 4 show that amplifications in all the regions 1q11–44 do not equally discriminate cluster 3. For example, rule

Rule 1: `Cluster3(X) ← 1q43-44(X) ∧ 1q12(X)`

characterizes cluster 3

best with a precision of 1. This means that amplifications in regions 1q43-44 and 1q12 characterizes cluster 3. It also covers 81 of the 88 samples in cluster 3. Nevertheless, amplifications in regions 1q11–44 shown in Figure 3 as discriminating regions, appear in at least one of the rules in the table on the left panel of the Figure 4 with varying degree of precision. Similarly, the second most discriminating rule for cluster 3 is: Rule 2: `Cluster3(X) ← 1q11(X)` which covers 78 positive samples and 9 negative samples.

The rules listed in the table on the left panel of Figure 4 also capture the multiresolution phenomenon in the data. We input only one resolution of data to the algorithm but the hierarchy of different resolutions is used as background knowledge. For example, the literal 1q43–44 denotes a joint region in coarse resolution thus showing that the algorithm produces results at different resolutions. The results at different resolutions improve the understandability and interpretability of the rules [8]. Furthermore, other information added to the background knowledge are amplification hotspots, fragile sites, cancer genes, which are discriminating features of cancers but do not show to discriminate any specific clusters present in the data. Therefore, such additional information would be better utilized in situations where the dataset contains not only cancer samples but also control samples which is unfortunately not the situation here as our dataset has only cancer cases.

3.4 Visualizing Semantic Rules and Clusters with Banded Matrices

The second way to use the exposed banded structure of the data is to display columns that were found to be important due to appearing in rules from Section 3.3. We achieve this by highlighting the chromosomal regions which appear in the rules. As shown in Figure 4, the highlighted band for cluster 1 spans chromosome regions 1q32–44. For cluster 3, the entire q–arm of the chromosome is highlighted, as indeed the instances in cluster 3 have amplifications throughout the entire arm. The regions 1q11–12 and 1q43–44 appear in rules with higher lift, in contrast to the other regions showing that the amplifications on the edges of the region are more important for the characterization of the cluster.

In summary, Figures 3 and 4 together offer an improved view of the underlying data. Figure 3 shows all the clusters on the data while Figure 4 shows only specific cluster and its associated rules. We achieve this by reordering the matrix rows by placing similar items closer together to form a banded structure [5], which allows easier visualization of the clusters and rules. It is important to reorder the rows independently of the clustering process. Because the reordering selected does not depend on the cluster structure discovered, the resulting figures offer new insight into both the data and the clustering.

4 Summary and Conclusions

We have presented a three-part data analysis methodology: clustering, semantic subgroup discovery, and pattern visualization. Pattern visualization takes advantage of the structure—in our case the bandedness of the matrix. The proposed

visualization allows us to explain the discovered patterns by combining different views of the data, which may be difficult to compare without a unifying visual display. In our experiments, we analyzed DNA copy number amplifications in the form of 0–1 data, where the clustering developed in previous work was augmented by explanatory rules derived from a semantic pattern mining approach combined by the facility to display the bandedness structure of the data.

The proposed semi–automated methodology provides complete analysis of a complex real-world multiresolution data. The results in the form of different clusters, rules, and visualizations are interpretable by the domain experts. Especially, the visualizations with banded matrix helps to understand the clusters and the rules generated by the semantic pattern mining algorithm. Furthermore, the use of the background knowledge enables us to analyze multiresolution data and garner results at different levels of multiresolution hierarchy. Similarly, the the obtained rules help to quantitatively prioritize chromosomal regions that are hallmarks of certain cancers among all the different chromosomal regions that are amplified in those cancer cases. In future work, we plan to extend the methodology and evaluate it using the wide variety of problems in comparison to some representative conventional methods.

Acknowledgement. This work was supported by Helsinki Doctoral Programme in Computer Science — Advanced Computing and Intelligent Systems (Hecse) and by the Slovenian Ministry of Higher Education, Science and Technology (grant number P-103), the Slovenian Research Agency (grant numbers PR-04431, PR-05540) and the SemDM project (Development and application of new semantic data mining methods in life sciences), (grant number J2–5478). Additionally, the work was supported by the Academy of Finland (grant number 258568), and European Commission through the Human Brain Project (Grant number 604102).

References

[1] Chen, C.-H., Hwu, H.-G., Jang, W.-J., Kao, C.-H., Tien, Y.-J., Tzeng, S., Wu, H.-M.: Matrix Visualization and Information Mining. In: Antoch, J. (ed.) COMPSTAT 2004 – Proceedings in Computational Statistics, pp. 85–100. Physica-Verlag HD (2004)

[2] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B (Methodological) 39(1), 1–38 (1977)

[3] Durkin, S.G., Glover, T.W.: Chromosome Fragile Sites. Annual Review of Genetics 41(1), 169–192 (2007)

[4] Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., Stratton, M.R.: A census of human cancer genes. Nature Reviews. Cancer 4(3), 177–183 (2004)

[5] Garriga, G.C., Junttila, E., Mannila, H.: Banded structure in binary matrices. Knowledge and Information Systems 28(1), 197–226 (2011)

[6] Hand, D., Mannila, H., Smyth, P.: Principles of Data Mining. Adaptive Computation and Machine Learning Series. MIT Press (2001)

[7] Hollmén, J., Seppänen, J.K., Mannila, H.: Mixture models and frequent sets: combining global and local methods for 0-1 data. In: Proceedings of the Third SIAM International Conference on Data Mining, pp. 289–293. Society of Industrial and Applied Mathematics (2003)
[8] Hollmén, J., Tikka, J.: Compact and understandable descriptions of mixtures of Bernoulli distributions. In: Berthold, M.R., Shawe-Taylor, J., Lavrač, N. (eds.) IDA 2007. LNCS, vol. 4723, pp. 1–12. Springer, Heidelberg (2007)
[9] Langohr, L., Podpecan, V., Petek, M., Mozetic, I., Gruden, K., Lavrač, N., Toivonen, H.: Contrasting Subgroup Discovery. The Computer Journal 56(3), 289–303 (2013)
[10] Lockwood, W.W., Chari, R., Coe, B.P., Girard, L., Macaulay, C., Lam, S., Gazdar, A.F., Minna, J.D., Lam, W.L.: DNA amplification is a ubiquitous mechanism of oncogene activation in lung and other cancers. Oncogene 27(33), 4615–4624 (2008)
[11] McLachlan, G.J., Peel, D.: Finite mixture models. Probability and Statistics – Applied Probability and Statistics, vol. 299. Wiley (2000)
[12] Myllykangas, S., Himberg, J., Böhling, T., Nagy, B., Hollmén, J., Knuutila, S.: DNA copy number amplification profiling of human neoplasms. Oncogene 25(55), 7324–7332 (2006)
[13] Myllykangas, S., Tikka, J., Böhling, T., Knuutila, S., Hollmén, J.: Classification of human cancers based on DNA copy number amplification modeling. BMC Medical Genomics 1(15) (May 2008)
[14] Novak, P., Lavrač, N., Webb, G.I.: Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining. Journal of Machine Learning Research 10, 377–403 (2009)
[15] Shaffer, L.G., Tommerup, N.: ISCN 2005: An Intl. System for Human Cytogenetic Nomenclature (2005) Recommendations of the Intl. Standing Committee on Human Cytogenetic Nomenclature. Karger (2005)
[16] Tikka, J., Hollmén, J., Myllykangas, S.: Mixture Modeling of DNA copy number amplification patterns in cancer. In: Sandoval, F., Prieto, A.G., Cabestany, J., Graña, M. (eds.) IWANN 2007. LNCS, vol. 4507, pp. 972–979. Springer, Heidelberg (2007)
[17] Trajkovski, I., Železný, F., Lavrač, N., Tolar, J.: Learning Relational Descriptions of Differentially Expressed Gene Groups. IEEE Transactions on Systems, Man, and Cybernetics, Part C 38(1), 16–25 (2008)
[18] Vavpetič, A., Lavrač, N.: Semantic Subgroup Discovery Systems and Workflows in the SDM-Toolkit. The Comput. J. 56(3), 304–320 (2013)
[19] Vavpetič, A., Novak, P.K., Grčar, M., Mozetič, I., Lavrač, N.: Semantic Data Mining of Financial News Articles. In: Fürnkranz, J., Hüllermeier, E., Higuchi, T. (eds.) DS 2013. LNCS, vol. 8140, pp. 294–307. Springer, Heidelberg (2013)
[20] Vavpetič, A., Podpečan, V., Lavrač, N.: Semantic subgroup explanations. Journal of Intelligent Information Systems (2013) (in press)
[21] zur Hausen, H.: The search for infectious causes of human cancers: Where and why. Virology 392(1), 1–10 (2009)