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.

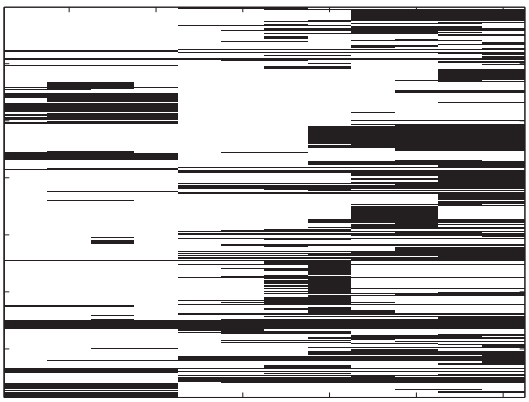

Figure 1: ***DNA copy number amplifications in chromosome-17, resolution 393.*** $\chi = (X_{ij})$, $X_{ij} \in \{0, 1\}$ **. Each row represents one sample of the amplification pattern for a patient and each column represents one of the chromosome bands.**

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.

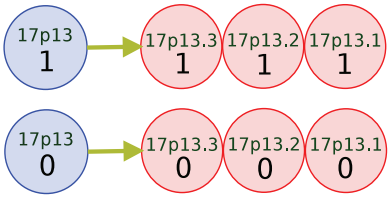

Figure 2: ***Schematic representation of upsampling where duplicate copies of similar cytogenetic bands are made in the higher resolution.***

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

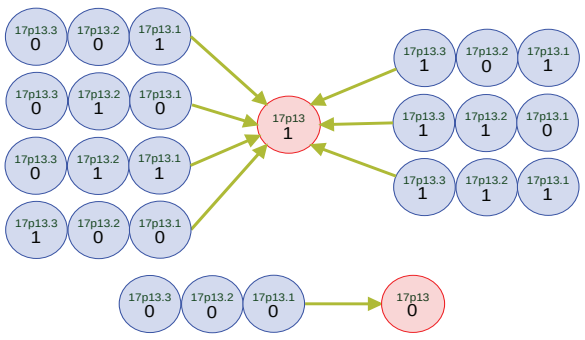

Figure 3: Schematic representation of OR-function downsampling procedure. Here the cytogenetic band in lower resolution is amplified if any of the bands in higher resolution is amplified. Cytogenetic band in lower resolution is not amplified only when none of the bands in higher resolution is amplified.

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.

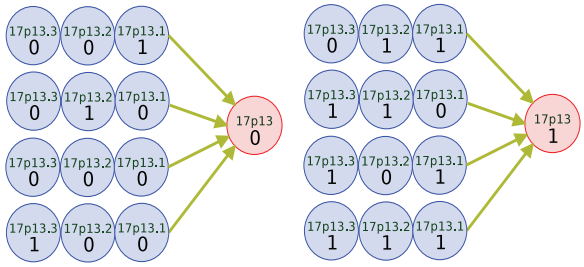

Figure 4: Schematic representation of majority decision downsampling procedure. Here the cytogenetic band in lower resolution is amplified if majority of the bands in higher resolution are amplified, otherwise it not amplified.

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 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.

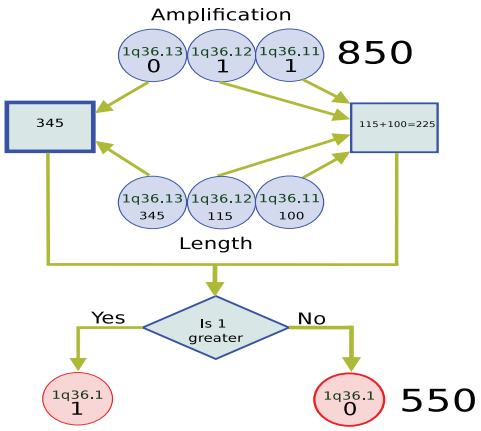

Figure 5: Schematic representation of weighted average downsampling procedure. Here the cytogenetic band in lower resolution is amplified if total length of the amplified bands in higher resolution is greater than the total length of unamplified bands, otherwise it not amplified. The figure is an example 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.

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

[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.

[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.

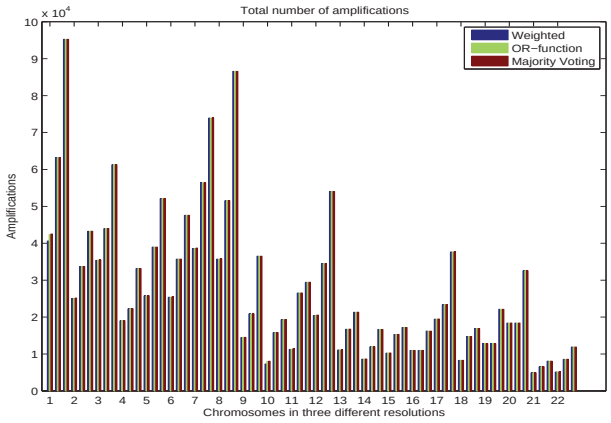

Figure 6: Total number of amplifications produced by the three different downsampling methods.

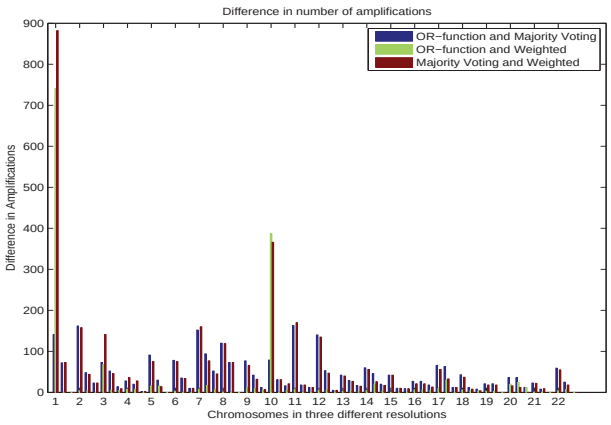

Figure 7: Difference in total number of amplifications produced by the three different downsampling methods.

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.

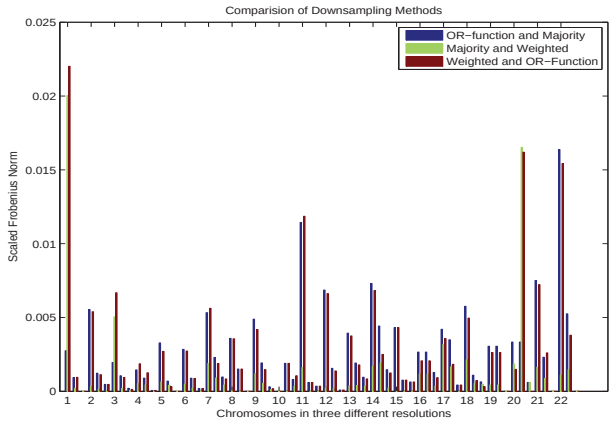

Figure 8: Comparison of three different downsampling methods : The difference measure used is scaled Frobenius norm.

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.

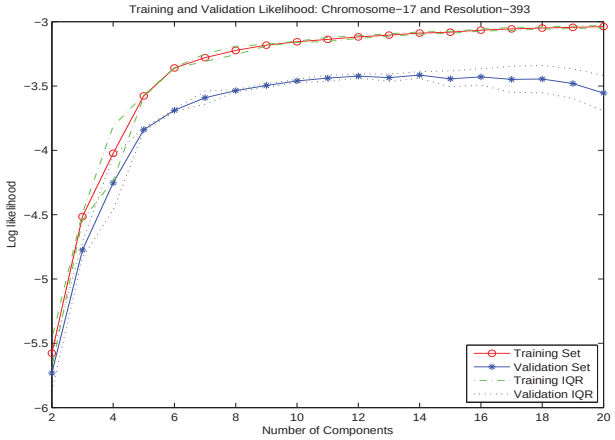

Figure 9: Model selection for the the original data in resolution 393. The averaged log-likelihood for training and validation sets in a 10-fold cross-validation setting for different number of components in chromosome-17: Resolution-393. The interquartile range(IQR) for 50 different training and validation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

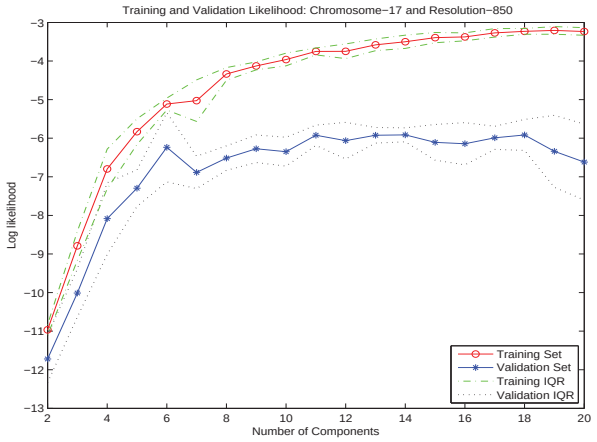

Figure 10: The averaged log-likelihood for training and validation sets in a 10-fold Cross validation setting for different number of components in chromosome17: Resolution 850. The interquartile range(IQR) for 10 different training and validation runs have also been plotted.

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.