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.
© Springer International Publishing Switzerland 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*).

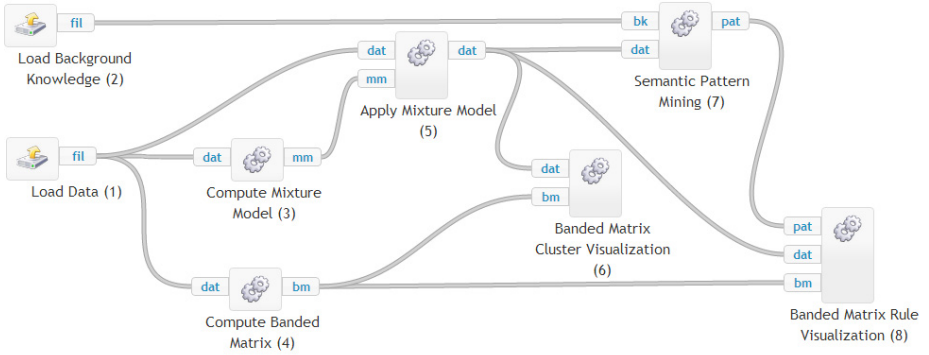

Fig. 1. Overview of the proposed three-part methodology

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) ← 1q43-44(X) ∧ 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.

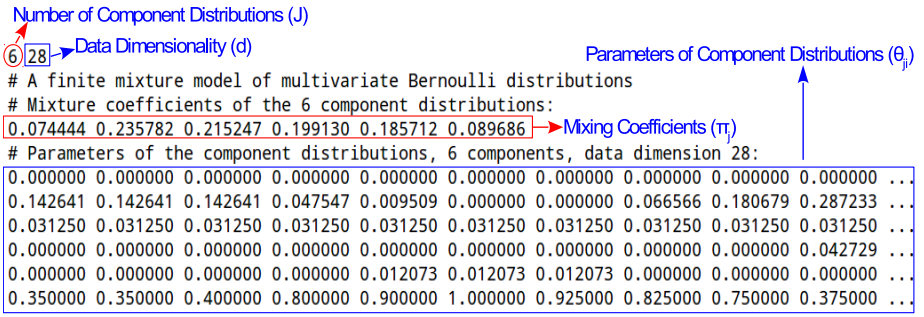

Fig. 2. Mixture model for chromosome 1. Figure shows first 10 dimensions of the total 28 for clarity.

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

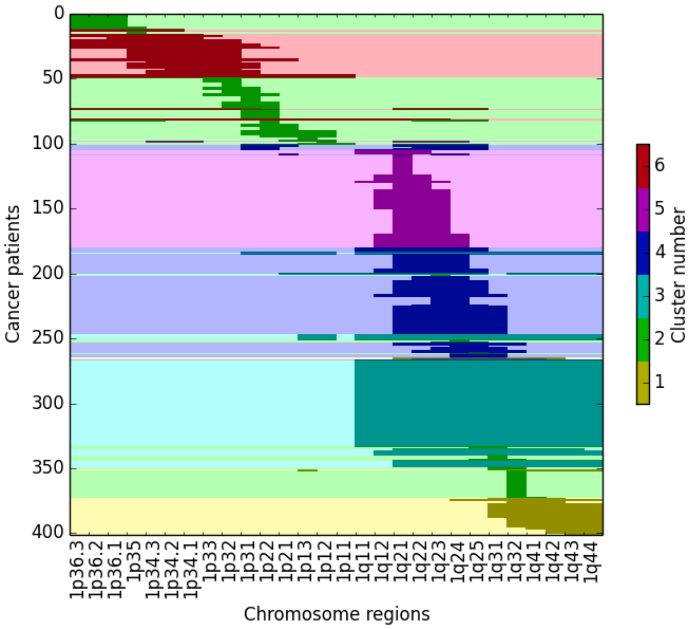

Fig. 3. Banded structure of the matrix with cluster information overlay

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.

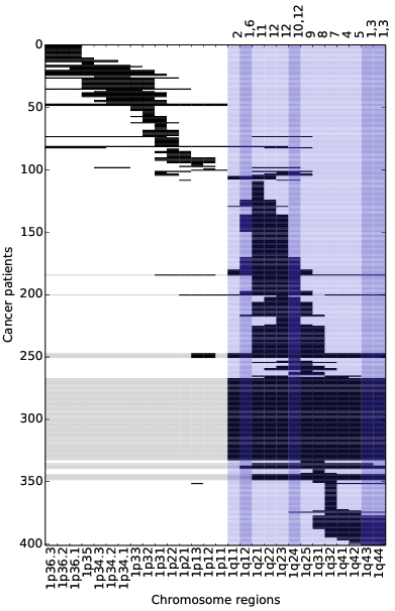

Fig. 4. Rules induced for cluster 3 (left) and visualizations of rules and columns for cluster 3 (right) with relevant columns highlighted. A highlighted column denotes that an amplification in the corresponding region characterizes the instances of the particular cluster. A darker hue means that the 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)