Compact and Understandable Descriptions of Mixtures of Bernoulli Distributions

Jaakko Hollmén and Jarkko Tikka

Helsinki Institute of Information Technology – HIIT
Helsinki University of Technology, Laboratory of Computer and Information Science, P.O. Box 5400, FI-02015 TKK, Espoo, Finland
Jaakko.Hollmen@tkk.fi, tikka@mail.cis.hut.fi

Abstract. Finite mixture models can be used in estimating complex, unknown probability distributions and also in clustering data. The parameters of the models form a complex representation and are not suitable for interpretation purposes as such. In this paper, we present a methodology to describe the finite mixture of multivariate Bernoulli distributions with a compact and understandable description. First, we cluster the data with the mixture model and subsequently extract the maximal frequent itemsets from the cluster-specific data sets. The mixture model is used to model the data set globally and the frequent itemsets model the marginal distributions of the partitioned data locally. We present the results in understandable terms that reflect the domain properties of the data. In our application of analyzing DNA copy number amplifications, the descriptions of amplification patterns are represented in nomenclature used in literature to report amplification patterns and generally used by domain experts in biology and medicine.

1 Introduction

In data analysis, the model should absorb the essentials about the data measured from a phenomenon and abstract away the irrelevant details about a particular data set. Parsimonious representations aim at particularly compact and simplified models. These kind of models offer an appealing basis for understanding and describing a phenomenon of interest. Previously, we have investigated parsimonious model representations in ecology [12], where we predicted nutrient concentrations in coniferous trees with a sparse regression methodology. In a time series prediction context, we have proposed a fast input selection method for long-term prediction [14] using a filter strategy. This strategy selects a possibly non-contiguous set of autoregressive variables with linear techniques and builds more complex non-linear prediction models using only the selected variables. In our experience, the parsimonious models are highly desired by domain experts, for instance, in biology, medicine, and ecology. In the models mentioned above, we have included roughly ten percent of the variables (in fact, parameters) compared to the models represented by all the parameters (full models). The sparse models still produce as accurate predictions as the full models. Another line of

M.R. Berthold, J. Shawe-Taylor, and N. Lavrač (Eds.): IDA 2007, LNCS 4723, pp. 1–12, 2007.

research where an attempt is made to concisely describe a data set is reported in [15]. We have presented a tool for automatically generating data survey reports for the modeler to be aware of the properties of the data set. While technically slightly different, the spirit still remains the same as in the current work: the focus is on describing the cluster structure and the contents of the clusters. The aim of the current paper is to present a way to summarize a finite mixture model for 0-1 data concisely and with a simple, domain-compatible representation.

Our research has been motivated by work in analyzing DNA copy number amplifications represented as 0-1 data by profiling [9] and by mixture modeling [13]. The mixture modeling approach offers an elegant way to model DNA amplification patterns in a probabilistic framework. However, the mixture models are summarized by arrays of numerical probability values that are hard to grasp. Therefore, we investigate how to describe the essential properties of the mixture models through the parameters of the models, or alternatively through the clustered data sets. Our proposed solution is based on the maximal frequent itemsets which are extracted from the clustered data sets. The descriptions are represented in the style of the descriptions used in literature to report amplification patterns and generally used by domain experts.

The rest of the paper is organized as follows: Sect. 2 describes the DNA copy number amplification data and our previous research in this context. Sect. 3 describes mixture models in the analysis of 0-1 data and the partitioning scheme for dividing the data in to cluster-specific data sets. The main topic of the paper — how to describe the mixture model for 0-1 data in a compact and understandable fashion — is explained in Sect. 4. Experiments are reported in Sect. 5 and the paper is summarized in Sect. 6. The nomenclature for the chromosome regions, which is used in the experimental part of the paper, is described in Appendix A.

2 DNA Copy Number Amplification Database

We have analyzed the database of DNA copy number amplifications collected with a bibliomics survey from 838 journal articles covering a publication period of ten years from 1992 until 2002 (for details, see [9]). DNA copy number amplifications are localized chromosomal aberrations that increase the number of copies of a chromosomal region from two to at least five. In the database, the DNA copy number amplifications are recorded for $N = 4590$ cancer patients in $d = 393$ chromosomal regions covering the whole human genome, and the observed data are the presence ($x_{ij} = 1$) or the absence ($x_{ij} = 0$) of DNA copy number amplifications for the patient i in the chromosomal region j, where $i = 1, \ldots, 4590$ and $j = 1, \ldots, 393$. For the case including only chromosome 1 presented later in the paper, the dimensions of the data are $N = 446$ and $d = 28$. The nomenclature for the chromosome regions used later in this paper is briefly described in Appendix A. In our previous work, we have analyzed a large 0-1 database of DNA copy number amplification patterns in human neoplasms [9]. We characterized the genome-wide data with cancer-specific amplification profiles with a probabilistic interpretation and

clustered the data with hierarchical clustering. Amplification-based clustering demonstrated that cancers with similar etiology, cell-of-origin or topographical location have a tendency to obtain convergent amplification profiles [9]. Furthermore, we applied independent component analysis (ICA) [7] to identify amplification hot spots, which are sparse, genome-wide factors defining statistically independent amplification sites in the data.

3 Mixture Models of DNA Copy Number Amplifications

Finite mixture models are widely used in data analysis and pattern recognition due to their flexibility and solid theoretical basis. The finite mixture model is parameterized by the number of components J and their corresponding mixing proportions π_j with non-negativity constraints $\pi_j \geq 0$ and necessary constraint $\sum_{j=1}^{J} \pi_j = 1$ so that the distribution integrates to one. Each of the component distributions is parameterized by a vector of parameters $\theta_j = (\theta_{j1}, \ldots, \theta_{jd})$. Considering the class of mixture models with J Bernoulli distribution components, each with dimensionality d, the model is summarized with $J + J \times d = J \times (1 + d)$ parameters and the appropriate conditional independence and independence assumptions. The probability of the data vector $\mathbf{x} = (x_1, \ldots, x_d)$ can be calculated as

$$P(\mathbf{x}) = \sum_{j=1}^{J} \pi_j \, P(\mathbf{x} \mid \theta_j) = \sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} \theta_{ji}^{x_i} (1 - \theta_{ji})^{1 - x_i}. \quad (1)$$

Learning from the data is most conventionally done in the framework of maximum likelihood estimation using the iterative Expectation-Maximization (EM) algorithm [3,16].

Recently, we modeled a refined version of the DNA copy number amplification database containing only malignant tumors, or cancers, and modeled the DNA copy number amplification patterns with finite mixtures of Bernoulli distributions in a chromosome-specific manner [13]. Model selection was performed for each chromosome in order to select an appropriate number of component distributions using a 5-fold cross-validation procedure repeated ten times. As a result, we got 23 mixture models for chromosomes $1, 2, \ldots, 22, X$ with altogether 111 component distributions. The Y chromosome was left out of the analysis due to lack of data. In our example model for the DNA copy number amplifications in chromosome 1, we have 6 component distributions and the data dimension $d = 28$, so we have $6 \times (1 + 28) = 174$ parameters. The resulting model (see Fig. 1) is determined in terms of the maximum likelihood estimates of the parameters, which might be hard to interpret since the parameters of the model may have different roles and do not obey any simple geometric similarity observable by the human eye. The subsequent problem is to consider the result set and think of the best possible way to convey the model to experts in biology and medicine. The details of the mixture modeling is reported in [13]. Here, we concentrate on presenting the models with a compact and understandable description. The description also has a direct relevance to diagnostic patterns of amplification that can be used by experts in their research or clinical work.

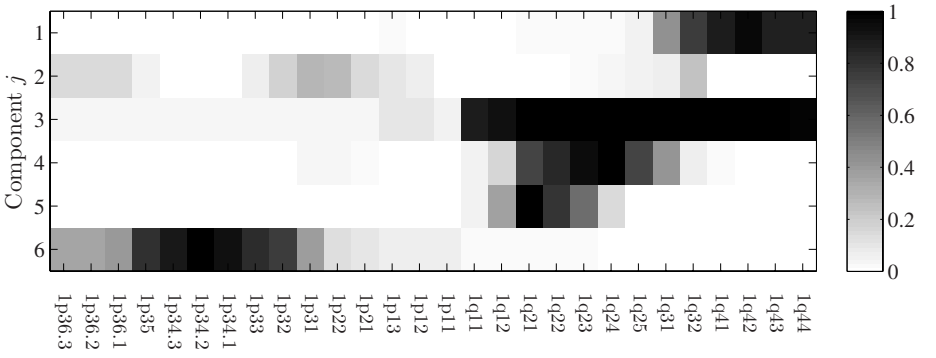

Fig. 1. Parameters $\theta_{ji}, j = 1, \ldots, 6,\ i = 1, \ldots, 28$ of the final mixture model. Each row represent the parameters of the corresponding component. The mixture proportions are $\pi_1 = 0.07, \pi_2 = 0.24, \pi_3 = 0.21, \pi_4 = 0.20, \pi_5 = 0.19$, and $\pi_6 = 0.09$. The names of the bands of the chromosome 1 (corresponding to 28 variables) are shown under the x-axis. For a brief explanation of the naming scheme of chromosomal regions, see Appendix A.

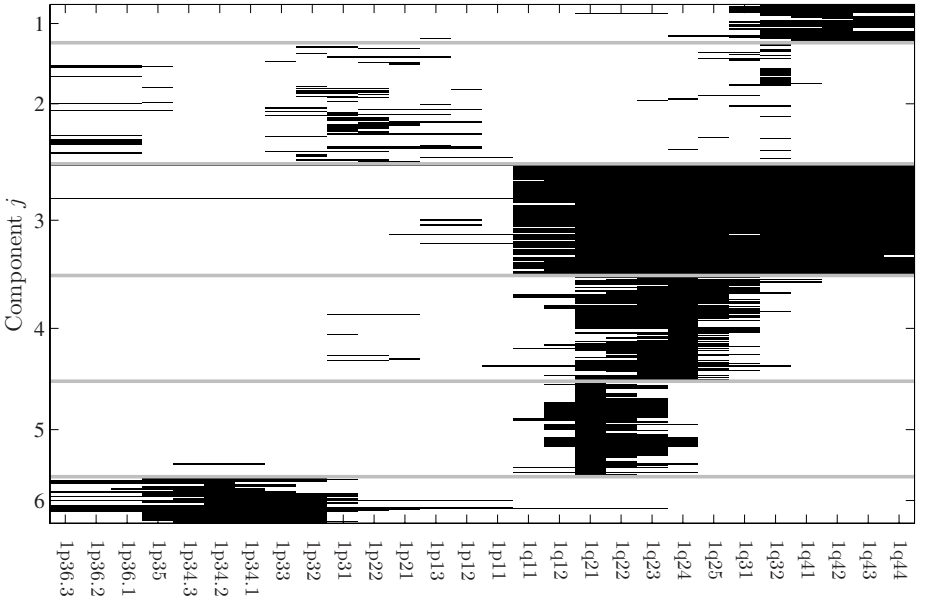

Fig. 2. The clustered DNA copy number amplification data is illustrated. Each horizontal line is one data vector; black areas mark ones and the white areas are zeroes in the data. The grey horizontal lines mark the partitions determined by the six probabilistic amplification patterns in the mixture model (see Fig. 1). The data in the clusters can easily be characterized, underlying the goodness of the clustering model and enabling compact and understandable cluster-specific descriptions of the model and the data.

The finite mixture model can be used in clustering by associating a cluster with each of the component distributions. Following the probabilistic approach, we can partition the original data set by allocating each data vector in to a cluster with the maximum posterior probability given the data vector. This maximum posterior allocation can be written with the help of Bayes's theorem as

$$P(j \mid \mathbf{x}) = \frac{P(j)P(\mathbf{x} \mid j)}{P(\mathbf{x})} = \frac{P(j)P(\mathbf{x} \mid j)}{\sum_{j'=1}^{J} P(j')P(\mathbf{x} \mid j')}. \tag{2}$$

Since the denominator in Eq. 2 is constant for all component distributions j, it is sufficient to partition the data vector $\mathbf{x}$ in to the cluster j^* using

$$j^* = \underset{j}{\operatorname{argmax}}\ P(j)P(\mathbf{x} \mid j) = \underset{j}{\operatorname{argmax}}\ \pi_j \prod_{i=1}^{d} \theta_{ji}^{x_i}(1-\theta_{ji})^{1-x_i}. \tag{3}$$

The clustering of the data set in to $J = 6$ partitions is illustrated in Fig. 2. In the following, we will consider the analysis of the clustered data.

4 Compact Description of the Mixture Models

We consider two alternative approaches to describe the multivariate Bernoulli distributions for modeling 0-1 data. The first approach is based on describing the model directly through the parameters of the mixture model (Sect. 4.1 and Sect. 4.2). The second approach to describe the model is with the help of the cluster-specific data sets that are achieved when the original data have been partitioned (Sect. 4.3). Our methodological contributions concentrate on the latter (describing the model though the data), but in the following, we also explore ways to characterize the model in terms of the parameters. It is also instructive to see how the methods are related to each other.

4.1 Modes of the Component Distributions

The simplest of all descriptions (and also the most compact) is to describe the model in terms of the modes of the respective component distributions. By definition, this only highlights the most probable elements of the data vectors described by the component distributions of the mixture model. If local chromosomal areas are sought for diagnostic purposes, this may be an ideal choice. However, this might suffer from the extremely simplified representation that is not faithful to the original data. For instance, if there are two distinct large probabilities, only one will be represented by the mode. For instance, component 2 of our example model in Fig. 1 can hardly be summarized by a mode, and neither can component 3, which has a large contiguous range of high probabilities.

4.2 Hypothetical Mean Organism

In bacteriology (see [5] and references), biochemical profiles of bacterial strain have been summarized by means of replacing probabilities θ_{ij} with their quantized counterparts a_{ij} as

$$a_{ij} = \begin{cases} 1 & \text{if } 1/2 \leq \theta_{ij} \leq 1, \\ 0 & \text{if } 0 \quad \leq \theta_{ij} < 1/2. \end{cases}$$

The profiles formed by the a_{ij} are called hypothetical mean organisms and abbreviated HMO. They represent the vectors of probabilities with their most likely realizations by quantizing each of the parameters to the most likely value of the each component of θ_j (either one or zero). One noteworthy aspect is that this could in fact be a non-existent case in the data set and more importantly, could be an impossible real-world situation. Nonetheless, it describes one realization that characterizes the parameters with the minimum quantization error.

4.3 Maximal Frequent Itemsets from Clustered Data

A conceptually different approach to describe the model is through the data sets that are created by partitioning the data by the maximum posterior rule in the hard clustering way (see Eq. 3). Then, any description of the cluster (and therefore the component distribution of that cluster), is calculated as a function of the clustered data sets. As mentioned before, the goal of the description is to be compact and understandable, so that domain experts can take full advantage of the modeling effort. Since our data is 0-1 data, any description must be specific to the nature of data. Frequent itemsets are one such description. Frequent itemsets are disjunctions of a set of attributes that co-occur in the data set.

As a typical example, imagine a market basket with possible items $X = \{\texttt{milk}, \texttt{butter}, \texttt{bread}\}$. If, for instance `milk` and `butter`, occur together in baskets with frequency of over 70 percent, we say that the $\{\texttt{milk}, \texttt{butter}\}$ itemset is 0.7-frequent. The Apriori algorithm for calculating the frequent itemsets is one of the classical algorithms [1,8] in data mining. However, if a market basket with items $\{\texttt{milk}, \texttt{butter}, \texttt{bread}\}$ is a frequent itemset with our pre-specified frequency threshold, all of its subsets are also frequent. This is the so-called anti-monotonicity property of frequent itemsets. For description purposes, this is somewhat wasteful, and certainly not compact. Instead, we turn our attention to maximal frequent itemsets [2], which are defined to be the largest (in cardinality) frequent itemsets that exceed the threshold. A set of maximal frequent itemsets is several orders of magnitude smaller than the usual set of frequent itemsets, making them ideal in providing compact descriptions of the items in the database [2]. In the above case, only the itemset $\{\texttt{milk}, \texttt{butter}, \texttt{bread}\}$ would be reported, and not any of its subsets. We have used an algorithm for finding maximal frequent itemsets called MAFIA [2]. The algorithm uses a search strategy that integrates a depth-first traversal of the itemset lattice with effective pruning mechanisms and it is especially efficient when the itemsets in the database are very long [2]. The need for mining maximal frequent itemsets becomes clear

when we look at the results, which are long indeed, and impractical to mine with the Apriori algorithm [1,8]. In our experiments, we have chosen to use the frequency threshold $\sigma = 0.5$, since then the extracted (maximal) frequent itemset would be present in the majority of cases in the cluster data. This could be further motivated by some majority voting protocol.

Previously, we have used the mixture of Bernoulli distributions in clustering 0-1 data in order to derive frequent itemsets from the cluster-specific data sets [6]. We investigated whether the frequent itemsets extracted in the clusters would be any different from the ones extracted globally. First, we clustered the data with a mixture of multivariate Bernoulli distributions and subsequently extracted frequent itemsets from the cluster-specific data sets. We introduced an L_1-norm based distance measure between sets of frequent itemsets and concluded that the frequent itemsets are markedly different from those extracted from the data set globally. On the basis of this previous work, we can expect different frequent itemsets to emerge from local, cluster-specific data sets and for them to describe the contents of the cluster more accurately. Frequent itemsets approximate in essence the marginal distribution of data, but can be also used as a basis in estimation of the joint distribution with the principle of maximum entropy [10,6].

The novelty in this paper compared to our previous work [6] is the real application that we are tackling, and we are trying to extract a compact and understandable description of the phenomenon. Furthermore, we are extracting descriptions of one model selected through a model selection procedure and not quantifying the differences over a range of models. Technically, we revert to maximal frequent itemsets and the translated descriptions based on those.

5 Experiments

We now describe the experimental work in training the mixture models from data and extracting maximal frequent itemsets from the cluster-specific data sets. The compact and understandable descriptions are based on the maximal frequent itemsets. The descriptions are presented in the original naming scheme for chromosomal regions. This nomenclature [11] is presented briefly in Appendix A. In our example chromosome 1, the chromosomal regions are coded with 28 regions based on the banding structure. The list of regions is: `1p36.3`, `1p36.2`, `1p36.1`, `1p35`, `1p34.3`, `1p34.2`, `1p34.1`, `1p33`, `1p32`, `1p31`, `1p22`, `1p21`, `1p13`, `1p12`, `1p11`, `1q11`, `1q12`, `1q21`, `1q22`, `1q23`, `1q24`, `1q25`, `1q31`, `1q32`, `1q41`, `1q42`, `1q43`, `1q44`. If we have an amplification of the three first chromosomal regions `1p36.3`, `1p36.2`, and `1p36.1` (denoted as a range `1p36.1`–`1p36.3`), the corresponding data vector would be $\mathbf{x} = (1, 1, 1, 0, \ldots, 0)$.

During the analysis of all chromosomes, the model selection procedure was used to select an appropriate complexity of the mixture model by varying the number of component distributions from $J = 2, \ldots, 20$. We trained 50 models by repeating 5-fold cross-validation ten times and selecting the model complexity with the largest validation log-likelihood. In one chromosome, this was seen to result in a model in which several component distributions were modeling

Table 1. Comparison of the different descriptions of the mixture of multivariate Bernoulli distributions for the chromosome 1 example. The parameters of the model are depicted in Fig. 1 and the clustered data in Fig. 2. The hypothetical mean organisms (HMO) are expressed in terms of the quantized probabilities a_{ij}. The modes of the component distributions are only given in cases when there is a clear single peak in the component distribution. Our compact descriptions (in the last column) are summarized as ranges from the contiguous maximal frequent itemsets extracted from the clustered data sets. The HMOs and our compact descriptions happen to describe the same chromosomal regions, although they are created using different paths.

j	HMO	mode	description based on itemsets
1	00000000000000000000000011111	1q42	1q32–1q44
2	00000000000000000000000000000	not given	none
3	00000000000000001111111111111	not given	1q11–1q44
4	00000000000000000011111000000	1q23	1q21–1q25
5	00000000000000000011100000000	1q21	1q21–1q23
6	00011111100000000000000000000	1p34.3	1p35–1p32

essentially the same data. In this case, a simpler model was selected by hand based on a hump of the validation likelihood. In all other chromosomes, a model selection procedure was followed.

5.1 Characterization of the Patterns

Repeating our proposed procedure for all chromosomes (except chromosome Y, which was discarded due to a low number of DNA copy number amplifications), we get in total 140 amplification patterns described by the maximal frequent itemsets with the frequency threshold $\sigma = 0.5$. The amplification patterns span a total of 337 chromosomal bands with some overlap between them. It is possible to have many itemsets for one component distribution; typically, they would be overlapping. The descriptions for the chromosome 1 example are listed in Table 1. To summarize the length distribution of all the itemsets, we extracted the following statistics: the average number of items in the itemsets is 5.1 and the median is 3. The 10th and 90th percentiles of the number of items are 1 and 12, respectively. The maximum number of items is 27. The large size of the individual itemsets underlines the impossibility of extracting frequent itemsets from the cluster-specific data: the number of all frequent itemsets (all subsets of the maximal frequent itemsets) would be about 146 million and would not be of any use in describing the models.

The mixture models seem to offer a nice way to model the domain with a probabilistic approach and with an excellent generalization ability, which is the ability to describe the general behavior of the amplification patterns. One characteristic of the identified models is their contiguous, smooth nature, which makes it very plausible for them to occur in nature. A surprising finding about the descriptions is that most of them describe contiguous chromosomal regions, that is,

spatially connected areas of the chromosome, although the extraction happens independently for the individual items. As the DNA copy number amplifications are expected to occur in the spatial manner [9], this speaks for the good quality of both the mixture model and the extracted descriptions.

As a reference, we extracted maximal frequent itemsets with a frequency threshold $\sigma = 0.5$ from all the data from chromosome 1 and got the following two itemsets: {1q21, 1q22} and {1q22, 1q23}. Another comparison would be to extract maximal frequent itemsets with $\sigma = 0.5/6$, the number 6 being the optimal number of clusters found with the aid of the model selection procedure. The resulting collection of itemsets was {1p31}, {1p22}, {1p36.1, 1p36.2, 1p36.3}, three overlapping itemsets with three items between 1p35 and 1p32, and a long itemset covering the whole 1q-arm. Two findings are striking: some spurious results emerge (1p36.1–1p36.3) and some results (1q-arm) shadow other interesting results found through partitioning of the data.

Another way to investigate the nature of the patterns is to compare them to some external data. One such data set that has close connections to DNA copy number amplifications is the fragile sites, which are discussed in more detail in our previous work [9]. There are 117 listed fragile sites in the genome (excluding the Y chromosome), of which 104 fragile sites map to our defined amplification patterns and 65 map to the ends of the amplification patterns. We would be interested to know if there is an unexpectedly large number of fragile sites in the ends of our amplification patterns, since this would indicate a possible explanation for DNA breakage associated with the amplification. We have compared the frequency of fragile sites in the ends of the amplification patterns and compared it with the frequency of fragile sites inside the patterns. A hypothesis test was executed with the help of a permutation test [4]. The frequency of fragile sites in the end regions is 0.3693 and inside the patterns it is 0.3086. Running the permutation test with 10000 repetitions, essentially sampling from the distribution of the null hypothesis by randomly picking 104 sites from the set of all possible sites, mapping them to the amplification patterns and calculating the frequency of border patterns and inside patterns, we get the differences for the random placements for the fragile sites. The p-value is calculated as a tail integral of the empirical distribution where the frequency of samples exceeds the true difference between the frequencies (one-tailed test). The resulting p-value is 0.0069, implying statistical significance of the findings. In absolute terms, the difference between the frequencies may not seem that large, but the relative difference is substantial. Thus, we claim scientifically relevant findings, as well.

6 Summary and Conclusions

Finite mixture models are a probabilistically sound and elegant way to model data, but can be hard to understand and describe compactly. In this paper, we have presented a way to describe the component distributions of the mixture models by describing the underlying cluster-specific data in terms of maximal frequent itemsets. The mixture model is used to model the whole data set as

a sum distribution in the global fashion and the frequent itemsets model the marginal distributions of the partitioned data in a local fashion; partitions coincide with the underlying structure of the data. In our case study in the analysis of DNA copy number amplification data, the cluster structure is well identified and the extracted maximal frequent itemsets summarize the marginal distributions in the clusters compactly. The descriptions are presented using the terminology used in literature, providing a compact and understandable summary of the models.

Acknowledgments

This work has been supported by the Academy of Finland in the Research Program SYSBIO (Systems Biology and Bioinformatics), grant number 207469. We thank Jouni Seppänen for insightful discussions on the methodological topics and Samuel Myllykangas for active and fruitful collaboration on cancer genomics. We also thank Paul Grouchy for comments on the manuscript.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the Twentieth International Conference on Very Large Data Bases (VLDB'94), pp. 487–499 (1994)
2. Burdick, D., Calimlim, M., Gehrke, J.: MAFIA: A maximal frequent itemset algorithm for transactional databases. In: Proceedings of the 17th International Conference on Data Engineering (ICDE'2001), pp. 443–452 (2001)
3. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B 39, 1–38 (1977)
4. Good, P.: Permutation Tests — A Practical Guide to Resampling Methods for Testing Hypotheses, 2nd edn. Springer, Heidelberg (2000)
5. Gyllenberg, M., Koski, T.: Probabilistic models for bacterial taxonomy. TUCS Technical Report No 325, Turku Centre of Computer Science (January 2000)
6. Hollmén, J., Seppänen, J.K., Mannila, H.: Mixture models and frequent sets: combining global and local methods for 0-1 data. In: Barbará, D., Kamath, C. (eds.) Proceedings of the Third SIAM International Conference on Data Mining, pp. 289–293. Society of Industrial and Applied Mathematics (2003)
7. Hyvärinen, A., Karhunen, J., Oja, E.: Independent component analysis. John Wiley & Sons, Chichester (2001)
8. Mannila, H., Toivonen, H., Verkamo, A.I.: Efficient algorithms for discovering association rules. In: Knowledge Discovery in databases: Papers form the AAAI-94 Workshop (KDD'94), pp. 181–192. AAAI Press (1994)
9. 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)
10. Pavlov, D., Mannila, H., Smyth, P.: Beyond independence: Probabilistic models for query approximation on binary transaction data. IEEE Transactions on Knowledge and Data Engineering 15(6), 1409–1421 (2003)

11. Schaffer, L.G., Tommerup, N. (eds.): An International System for Human Cytogenetic Nomenclature. S. Karger, Basel (2005)
12. Sulkava, M., Tikka, J., Hollmén, J.: Sparse regression for analyzing the development of foliar nutrient concentrations in coniferous trees. Ecological Modeling 191(1), 118–130 (2006)
13. Tikka, J., Hollmén, J., Myllykangas, S.: Mixture modeling of DNA copy number amplification patterns in cancer. In: Sandoval, F., Prieto, A., Cabestany, J., Graña, M. (eds.) IWANN 2007. LNCS, vol. 4507, pp. 972–979. Springer, Heidelberg (2007)
14. Tikka, J., Lendasse, A., Hollmén, J.: Analysis of fast input selection: Application in time series prediction. In: Kollias, S., Stafylopatis, A., Duch, W., Oja, E. (eds.) ICANN 2006. LNCS, vol. 4132, pp. 161–170. Springer, Heidelberg (2006)
15. Vesanto, J., Hollmén, J.: An automated report generation tool for the data understanding phase. In: Abraham, A., Koeppen, M. (eds.) Proceedings of the First International Workshop on Hybrid Intelligent Systems (HIS'01), pp. 611–625. Springer, Heidelberg (2002)
16. Wolfe, J.W.: Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research 5, 329–350 (1970)

A Nomenclature for the Chromosomal Regions

The naming scheme for chromosomal regions will be briefly presented in order to understand the proposed description scheme of mixture models for 0-1 data in our case of DNA copy number amplifications. In essence, the terminology is used by domain experts in literature when addressing chromosomal regions. Idiograms of G-banding patterns for normal human chromosomes at five different levels of resolution are presented in [11].

Using the resolution of the data used in this paper, the whole human genome is divided in to 393 chromosomal regions, each with its own systematic name. The example chromosome, chromosome 1, is divided in to 28 regions. The nomenclature follows an irregular, hierarchical naming scheme, where each region may (or may not) be divided in to smaller subregions. The whole chromosome is denoted by `1`, which consists of two chromosome arms, the shorter arm named `1p` and the longer one named `1q`. The next levels will add numbers subsequent to `1p` and `1q`, as for instance in `1q12`. The name `1q12` tells us that the region belongs to chromosome `1`, arm `q`, and region `12`. The `1q12` is not divided any further in our resolution. Another example is the regions `1q21.1`, `1q21.2`, `1q21.3`, which are names of regions on a finer scale and where the part after the decimal period denotes sub-regions of the chromosome region `1q21`. Which regions are and which are not divided to finer subparts is determined by the resolution of the naming scheme and the properties of the genome. Chromosomal regions are often expressed as continuous ranges as for instance `1q21`–`1q23`.

The whole list of regions in the chromosome 1 is as follows: `1p36.3`, `1p36.2`, `1p36.1`, `1p35`, `1p34.3`, `1p34.2`, `1p34.1`, `1p33`, `1p32`, `1p31`, `1p22`, `1p21`, `1p13`, `1p12`, `1p11`, `1q11`, `1q12`, `1q21`, `1q22`, `1q23`, `1q24`, `1q25`, `1q31`, `1q32`, `1q41`, `1q42`, `1q43`, `1q44`. The DNA copy number amplification data expressed in terms of

chromosomal regions can be transformed to 0-1 data according to the amplification status of the corresponding regions $\mathbf{x} = (x_1, x_2, x_3, \ldots, x_{27}, x_{28}) = (x_{\texttt{1p36.3}}, x_{\texttt{1p36.2}}, x_{\texttt{1p36.1}}, \ldots, x_{\texttt{1q43}}, x_{\texttt{1q44}})$. The 0-1 data is used in the modeling; the proposed methodology produces compact descriptions of the models in the terminology originally used by domain experts.