Chapter 15

From Data to Knowledge Research Unit

Heikki Mannila, Jaakko Hollmén, Kai Puolamäki, Ella Bingham, Johan Himberg, Robert Gwadera, Hannes Heikinheimo, Antti Ukkonen, Jouni K. Seppänen, Nikolaj Tatti, Heli Hiisilä, Antti Rasinen, Mikko Korpela, Janne Toivola

15.1 Data mining at the Pattern Discovery group

The Pattern Discovery group in Otaniemi concentrates on combinations of pattern discovery and probabilistic modeling in data mining: pattern discovery aims at finding local phenomena, while modeling often aims at global analysis.

Pattern discovery techniques can be very efficient in finding frequently occurring patterns from large masses of data. Techniques for this task include both algorithmics in the traditional computer science sense and probabilistic methods.

The research topics include research on frequent itemsets, latent variable models and the problem of finding ordering from the data. The application areas include gene expression data, and paleontological and ecological data analysis.

The Pattern Discovery group in the Laboratory of Computer and Information Science is part of the From Data to Knowledge research unit, which is located in part in the Department of Computer Science at the University of Helsinki.

15.2 Discovering orderings

Hannes Heikinheimo, Heikki Mannila, Kai Puolamäki, Antti Ukkonen

In many applications, such as paleontology, medical genetics and ranking of user preferences, the 0-1 data has an underlying unknown order. In the case of paleontology, for example, we have a collection of fossil sites, which can be loosely regarded as a snapshot of the set of taxa that lived at a certain location at approximately same time. Sites and their taxa may be described as an 0-1 occurence matrix, like the one shown in figure 15.2, where the rows correspond to sites and the columns correspond to the taxa: one in entry (i, j) means that fossilized remnants of taxon j has been found at site i.

In paleontology seriation, or temporal ordering of fossil sites, is a central and difficult problem. The geological age determination is inaccurate, and some finds are classified inaccurately and many finds are missing. The underlying temporal order is partial; for example, two fossil sites may be from the same paleontological time period and hence this pair of sites has no preferred temporal order.

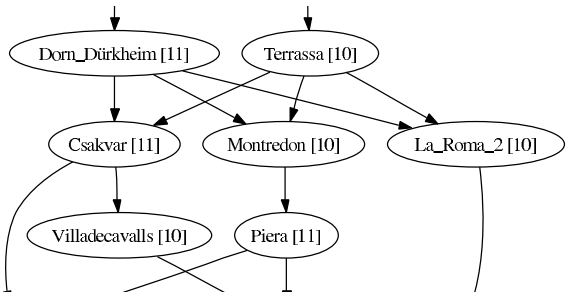

Figure 15.1: A subset of the discovered partial temporal order for 124 fossil sites, obtained by analyzing the occurence data on large late Cenzoic mammals.

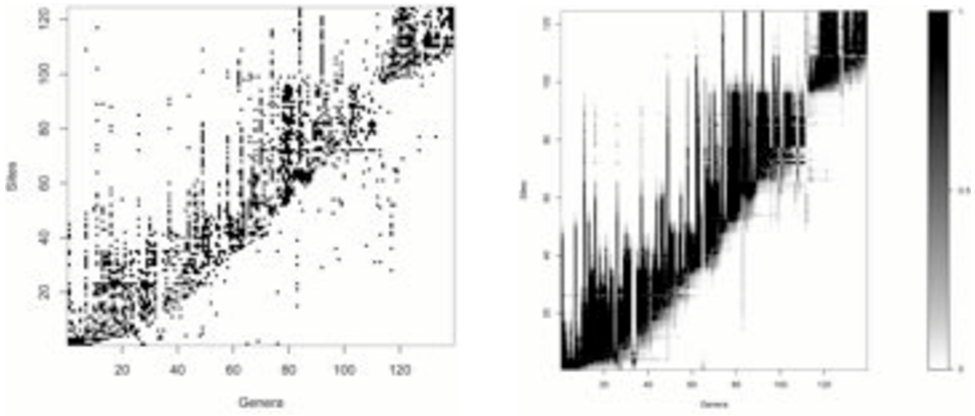

Figure 15.2: The figure on the left shows the original paleontological data in the preferred temporal order, given by the MCMC algorithm, black points denoting paleontological finds (ones in the occurence matrix). The figure on the right shows the probabilities that the genera were alive during the period of a find, given by our probabilistic model.

We have developed methods that can be used to find partial orders, based on the 0-1 occurence matrices. One approach is to study fragments of order, i.e., small subsets of

items to be ordered, and find the preferred total order for the fragment by minimizing a score function. [3] In case of the paleontological data, a score function for a given ordering of sites can be for example the number of changes from 1 to 0 for a given genus. The fragments can then be used to construct a partial order. like one shown in figure 15.1.

We have also developed probabilistic models to solve the ordering problem. We have solved the models with Markov Chain Monte Carlo (MCMC) method, which is able to — unlike earlier methods used to analyze the seriation problem — actually produce a fine-grained temporal ordering. [2] Our method has also been used to identify false finds the fossil databases, and also genera with unusual ecological characteristic.

Another application of partial orders we are working on is the description and summarization of large sets of total orders (rankings). [3] Given several total orders of a set of items, it is possible to determine one or a few partial orders that describe the original total orders well.

In addition to the fossil data, our group is in the process of analyzing other large ecological datasets. [1] Via cluster analysis we have already gained some interesting and ecologically relevant results on spatial distributions of mammalian metacommunities in Europe. Furthermore, this study has risen theoretically interesting questions for further methodological studies on cluster analysis in the context of data with prevailing spatial relationships.

References

[1] Hannes Heikinheimo. Inferring taxonomic hierarchies from 0-1 data. Master's thesis, Helsinki University of Technology, 2005.

[2] Kai Puolamäki, Mikael Fortelius, and Heikki Mannila. Seriation in paleontological data using Markov Chain Monte Carlo methods. *PLoS Computational Biology*, 2(2):e6, February 2006. http://dx.doi.org/10.1371/journal.pcbi.0020006.

[3] Antti Ukkonen. Data mining techniques for discovering partial orders. Master's thesis, Helsinki University of Technology, 2004.

[4] Antti Ukkonen, Mikael Fortelius, and Heikki Mannila. Finding partial orders from unordered 0-1 data. In *Proceedings of the 11th ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2005.

15.3 Theoretical aspects of data mining

Jaakko Hollmén, Heikki Mannila, Jouni K. Seppänen, Nikolaj Tatti

Nowadays one often encounters high-dimensional data in practical applications. Thus methods and algorithms to analyse such data are highly needed. A typical example of such data is 0–1 data, i.e., the case when the data consists of vectors whose elements are 0 and 1. Although it seems that this kind of data is the simplest one, many applications include analysis of such data. For example, binary data can be generated from text documents such that each element of a binary vector represent some particular word, set to 1 if this word is present in a document and 0 otherwise. Different databases provide a large and important class of applications. For example, in market basket data each vector represent a transaction and the elements represent different products. More such examples can be obtained from course participation data or citation data. Binary data can be also obtained from genome data, e.g., single nucleotide polymorphisms (SNP) are a direct example of this.

Frequent itemsets are one of the best known concepts in 0–1 data mining: an itemset is frequent in a database if its items co-occur in sufficiently many records. Since the inception of frequent itemset mining as a solution to the association rule mining problem in 1996, several methods for finding all frequent itemsets have been proposed, and algorithms for this task continue to be a large research area within data mining. A question that we feel has not been satisfactorily addressed is that of using the itemsets: what do they tell us about the original data?

One way to answer this question is to use frequent itemsets for query approximation. Given a Boolean query ϕ over the attributes of the original data, how good approximations can one obtain using only the frequent itemsets? An answer that is in principle complete was given in [1]: the itemsets can be seen as the conditions of a linear program, whose objective function can be minimized or maximized to find the minimum and maximum of the Boolean query. This solution was shown by experiments to be useful in some cases, but it has the intrinsic problem that the size of the linear program is exponential in the number of variables. In fact, we show in [4] that such query problems are NP-complete. However, we show in [5] that under some assumptions we can drastically reduce the number of variables and thus ease the computational burden.

Another possibility is to use a combinatorial algorithm to approximate the query: for example, if the query is a disjunction of attributes,

$$\phi = A_1 \vee A_2 \vee \cdots \vee A_k,$$

its answer is an inclusion-exclusion sum

$$\sum_{j=1}^{k} f(A_j) - \sum_{1 \le i < j \le k} f(A_i A_j) + \sum_{1 \le h < i < j \le k} f(A_h A_i A_j) + \cdots + (-1)^{k+1} f(A_1 A_2 \cdots A_k),$$

where f denotes the frequency of an itemset. Now if only the frequencies of frequent itemsets are filled in this sum, how far can it be from the correct result? The answer is twofold: in theory, the worst-case bound for the algorithm is very large, and a construction exists that shows the bound to be tight; but in practice, the approximations tend to be much closer to the correct answer than in the worst case. The theoretical part was addressed in [3], where the above approach was also extended to arbitrary Boolean formulas from the simple disjunctions. The practical results are as yet unpublished.

Finally, we comment in [2] on the recent idea of hypercube segmentation by Jon Kleinberg et al. Hypercube segmentation is one formalization of clustering 0–1 data. We show that the analysis by Kleinberg et al. of their algorithm is nearly tight, and if the approximation guarantee is to be significantly improved, the approach of selecting data vectors as cluster centers is not sufficient.

References

[1] Artur Bykowski, Jouni K. Seppänen, and Jaakko Hollmén. Model-independent bounding of the supports of boolean formulae in binary data. In Rosa Meo, Pier Luca Lanzi, and Mika Klemettinen, editors, *Database Support for Data Mining Applications: Discovering Knowledge with Inductive Queries*, volume 2682 of *Lecture Notes in Artificial Intelligence*, pages 234–249. Springer-Verlag, 2004.

[2] Jouni K. Seppänen. Upper bound for the approximation ratio of a class of hypercube segmentation algorithms. *Information Processing Letters*, 93(3):139–141, February 2005.

[3] Jouni K. Seppänen and Heikki Mannila. Boolean formulas and frequent sets. In Jean-François Boulicaut, Luc de Raedt, and Heikki Mannila, editors, *Constraint-based mining and inductive databases*, volume 3848 of *Lecture Notes in Artificial Intelligence*. Springer, 2005. To appear.

[4] Nikolaj Tatti. Computational complexity of queries based on itemsets. *Information Processing Letters*. In press.

[5] Nikolaj Tatti. Safe projections of binary data sets. *Acta Informatica*. In press. doi:10.1007/s00236-006-0009-9

15.4 Extending frequent itemsets: dense itemsets and tiles

Heikki Mannila, Jouni K. Seppänen

Another question related to frequent itemsets concerns extending their definition to relax the requirement of perfect co-occurrence: highly correlated items may form an interesting set, even if they never co-occur in a single record. The problem is to formalize this idea in a way that still admits efficient mining algorithms. Dense itemsets [2] are defined in a manner similar to frequent itemsets and can be found using a similar algorithm.

Another way to approach finding non-perfectly co-occurring items was defined in [1] and named "tiles". A spectral algorithm was used to rearrange the data matrix so that interesting sets of items become contiguous in both dimensions, and then these contiguous regions were found using a local search algorithm. This solution can also find non-perfectly anti-co-occurring items and hierarchical models where smaller tiles are used as exceptions to larger ones. An example of finding a hierarchical tile model is shown in Figure 15.3, where the leftmost pane shows the original data, the middle one shows the result of reordering, and the rightmost pane shows a model consisting of ten tiles.

An underlying theme connecting these topics is the interplay of two data mining objectives, local patterns and descriptive models. Frequent itemsets are an example of patterns: interesting phenomena occurring in some small part of the data. In contrast, a descriptive model tells us something interesting about the whole data. The query approximation problem is motivated by a desire to convert a frequent itemset collection into a model that can be used to answer queries about the data. Dense itemsets can be used to create a description of the data using e.g. a greedy algorithm on the dense itemsets. While frequent itemsets could be used similarly, the requirement of complete co-occurrence hinders the effort, and with dense itemsets the results are more interesting. Tile models are similarly descriptive models built from local patterns.

References

[1] Aristides Gionis, Heikki Mannila, and Jouni K. Seppänen. Geometric and combinatorial tiles in 0-1 data. In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, editors, *Knowledge Discovery in Databases: PKDD 2004*, volume 3202 of *Lecture Notes in Artificial Intelligence*, pages 173–184. Springer, 2004.

[2] Jouni K. Seppänen and Heikki Mannila. Dense itemsets. In Ronny Kohavi, Johannes Gehrke, William DuMouchel, and Joydeep Ghosh, editors, *Proceedings of the Tenth*

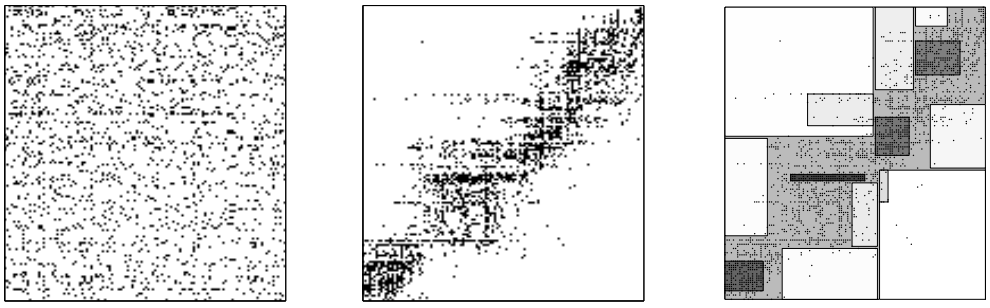

Figure 15.3: Example data first reordered, then hierarchically tiled

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), pages 683–688. ACM, 2004.

15.5 Basis segmentation

Ella Bingham, Heli Hiisilä, Heikki Mannila

In this project we have combined two techniques of multidimensional data analysis: segmentation of sequences, and dimensionality reduction. Both techniques reduce the complexity of representing the data. We use these techniques together, by finding the segment boundaries in a way that utilizes a low rank representation of the data.

We propose three different algorithms for the task, All of them consist of existing methods of segmentation (namely, optimal segmentation by dynamical programming) and dimensionality reduction (namely, principal component analysis which is optimal in the mean-squared-error sense).

- Algorithm SEG-PCA first segments the data, and then reduces the dimensionality of the data consisting of the segment means.
- Algorithm SEG-PCA-DP first segments the data, then reduces the dimensionality of the data consisting of the segment means, and then refines the segmentation by using the basis vectors of the dimensionality reduction.
- Algorithm PCA-SEG first decreases the dimensionality, and then segments the lower dimensional data.

As an application we might consider financial time series such as exchange rates of currencies. Each currency constitutes one dimension in the multidimensional data. Segmenting corresponds to splitting the time into different economical phases, and dimensionality reduction corresponds to finding dependencies between different currencies.

In a forthcoming paper [1] we have demonstrated our algorithms on exchange rate data, DNA sequences and meteorological time series.

References

[1] Ella Bingham, Aristides Gionis, Niina Haiminen, Heli Hiisilä, Heikki Mannila, Evimaria Terzi. Segmentation and dimensionality reduction. *2006 SIAM Conference on Data Mining, April 20-22, 2006, Bethesda, Maryland, USA.*

15.6 Data mining in bioinformatics

Ella Bingham, Heli Hiisilä, Johan Himberg, Jaakko Hollmén, Mikko Korpela, Heikki Mannila, Antti Rasinen, Jouni Seppänen, Janne Toivola

Bioinformatics is a new collaborative area of science that has risen out of the need to involve computer scientists in the analysis of data-intensive problems in biology and medicine. Data analysis plays an important role in gene expression studies, where high-dimensional, noisy microarray measurement matrices have to be processed to yield meaningful biological knowledge. This knowledge is needed in order to understand the functions of the genome on the whole and the role played by individual genes in particular diseases. This helps in developing diagnostic tools for early detection of diseases such as cancer [6], for instance. In addition to microarray measurememements, auxiliary data set are helpful in reducing the uncertainty in the analysis. Auxiliary data sets include additional gene expression measurements using independent measurement platforms for validation, comparative genomic hybridization to measure gene copy number changes, tissue microarrays, and publicly available databases of gene expression and annotation databses, such as the gene ontology databases. Integration of different data sets is thus seen as an important aspect of data mining in bionformatics [2, 1].

In collaboration with researchers from the University of Helsinki and the Occupational Health Institute of Finland, we are involved in various projects involving cancer [4, 5, 6], also in analyzing cancer patients with work-related asbestos-exposure. Asbestos is a well known lung cancer causing mineral fiber.

Amplification profiling of human neoplasms In a recent study, we analyzed cancer-related gene amplification patterns collected from published literature. The data was collected at chromosome band-specific resolution from 838 published chromosomal comparative genomics hybridization studies for more than 4500 cases. We identified type-specific amplification profiles for each of the 73 cancer types (Fig. 15.4). Furthermore, relationships between the cancer types can be analyzed by a clustering solution relating the profiles with a similarity measure and performing a hierarchical clustering. In order to reveal generally interesting amplification patterns for cancer in general, we have identified amplification hot spots by means of independent component analysis. These genome-wide, sparse patterns of amplification offers an avenue for further exploration of genomic alterations of cancer.

Dependencies between transcription factor binding sites Gene expression of eucaryotes is regulated through transcription factors which are molecules able to attach to the binding sites in the DNA sequence. These binding sites are small pieces of DNA usually found upstream from the gene they regulate. As the binding sites play an important role in the gene expression, it is of interest to find out their characteristics.

In this project we look for dependencies and independencies between these binding sites using independent component analysis, non-negative matrix factorization, probabilistic latent semantic analysis and the method of frequent sets. The data used are human gene upstream regions and possible binding sites listed in a biological database. Also, data from the baker's yeast genome is analyzed. The results of the project are described in [3].

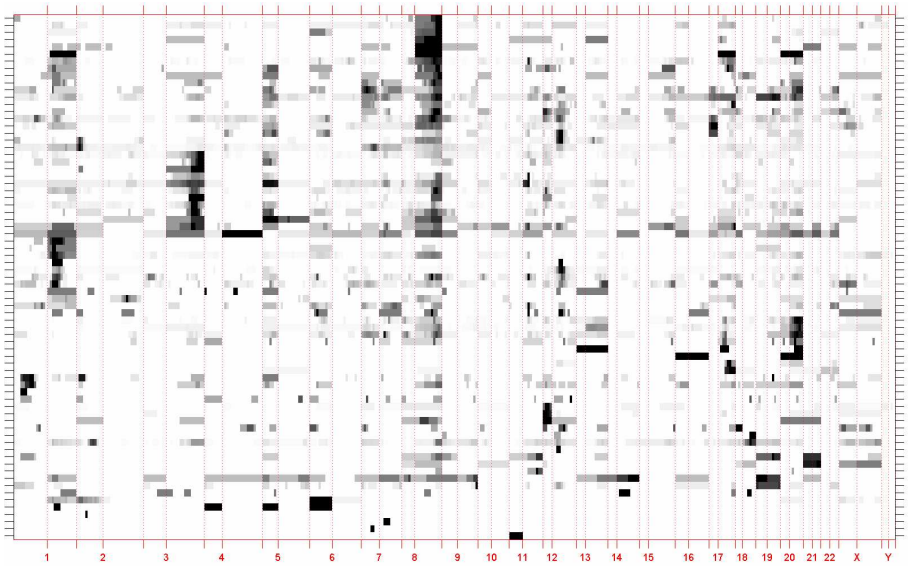

Figure 15.4: The cancer specific DNA amplification profiles are illustrated. The chromosomes are plotted on the horizontal axis one after the other in increasing order (starting with chromosome 1 on the left part of the figure). The profiles representing 73 cancer type profiles are plotted as rows.

References

[1] Catherine Bounsaythip, Erno Lindfors, Peddinti V. Gopalacharyulu, Jaakko Hollmén, and Matej Orešič. Network-based representation of biological data for enabling context-based mining. In *Proceedings of KRBIO'05, International Symposium of the Knowledge Representation in Bioinformatics*, pages 1–6, June 2005.

[2] Peddinti V. Gopalacharyulu, Erno Lindfors, Catherine Bounsaythip, Teemu Kivioja, Laxman Yetukuri, Jaakko Hollmén, and Matej Orešič. Data integration and visualization system for enabling conceptual biology. *Bioinformatics*, 21(Suppl.1):i177–i185, 2005.

[3] Heli Hiisilä and Ella Bingham. Dependencies between transcription factor binding sites: Comparison between ICA, NMF, PLSA and frequent sets. *Proceedings of the 4th IEEE International Conference on Data Mining, November 1-4, 2004, Brighton, UK*, pp. 114–121, 2004.

[4] Eeva Kettunen, Sisko Anttila, Jouni K. Sepp¡E4¿nen, Antti Karjalainen, Henrik Edgren, Irmeli Lindstr¡F6¿m, Reijo Salovaara, Anna-Maria Niss¡E9¿n, Jarmo Salo, Karin Mattson, Jaakko Hollmén, Sakari Knuutila, and Harriet Wikman. Differentially expressed genes in nonsmall cell lung cancer: expression profiling of cancer-related genes in squamous cell lung cancer. *Cancer Genetics and Cytogenetics*, 149(2):98–106, 2004.

[5] E. Kettunen, A.G. Nicholson, B. Nagy, J.K. Sepp¡E4¿nen, T. Ollikainen, G. Ladas, V. Kinnula, M. Dusmet, S. Nordling, J. Hollmén, D. Kamel, P. Goldstraw, and S. Knuutila. L1CAM, INP10, P-cadherin, tPA and ITGB4 over-expression in malignant pleural mesotheliomas revealed by combined use of cDNA and tissue microarray. *Carcinogenesis*, 26(1):17–25, 2005.

[6] Harriet Wikman, Jouni K. Sepp¡E4¿nen, Virinder K. Sarhadi, Eeva Kettunen, Kaisa Salmenkivi, Eeva Kuosma, Katri Vainio-Siukola, Balint Nagy, Antti Karjalainen, Thanos Sioris, Jarmo Salo, Jaakko Hollmén, Sakari Knuutila, and Sisko Anttila. Caveolins as tumor markers in lung cancer detected by combined use of cDNA and tissue microarrays. *Journal of Pathology*, 203:584–593, 2004.