Chapter 19

From Data to Knowledge Research Unit

Heikki Mannila, Jaakko Hollmén, Jouni K. Seppänen, Kalle Korpiaho, Johanna Tikanmäki, Ella Bingham

19.1 Data mining: discrete pattern discovery and probabilistic techniques

The pattern discovery group works on different techniques for the discovery of discrete patterns and on methods for finding various types of probabilistic models for discrete data. An overview of some of the methods was published in 2001 [1].

Modeling of sequences: partial orders, components, and random projections

An example of the work is the article [2]. The problem considered is finding good descriptions of data sets consisting of sequences of discrete values. The chosen model class consists of partial orders: a partial order can be considered to describe all total orders (i.e., sequences) that are compatible with it. Thus the task is to find a collection of partial orders that describes as large a fraction of the sequences in the input data as possible. In addition to this requirement, the partial orders should be as specific as possible, i.e., the number of all total orders compatible with the partial orders should be small. Computing the number of all total orders that are compatible with a partial order is a $\sharp$P-complete problem, so the model class is restricted to the set of series-parallel partial orders.

These requirements specify an optimization problem of finding a set of partial orders maximizing a scoring function. In [2] an algorithm is given for this problem, and empirical results are reported.

Long sequences of data items arise in many applications, such as telecommunications network management. One of the problems there is locating whether a recent combination of events has appeared before. In [3] we considered the use of random projections in finding similar situations. It turns out that the basic random projection approach can be profitably applied also in this situation.

Another problem in the analysis of sequences is finding whether a sequence of discrete symbols can be considered to be the independent mixture of components. That is, the task is to find a discrete analogue of independent component analysis (ICA). In [4] we considered ways of defining this notion and gave an algorithm for this task.

A sequence of discrete events can be described by giving the intensity of the events, assuming that each event is generated by a Poisson process with piecewise constant intensity. In [5] we use Markov chain Monte Carlo to obtain posterior distributions on the intensity. While this method is flexible, it is computationally very intensive. We considered also approaches based on dynamic programming, and showed that careful pruning of the candidate space of change points of the intensity functions can be used to obtain very fast methods for finding nearly optimal descriptions. Similar techniques are used in time series segmentation for the purpose of context recognition in mobile phones in [6].

Large-dimensional discrete data

Various applications, such as market basket data analysis and document analysis, result in large collections of high-dimensional count data. Such data sets are typically sparse: most entries are zero. The basic fast counting technique for such data sets is the finding of frequent sets and association rules. While there are lots of algorithms for this task, actual use of the methods is rare, as the resulting summaries of the data are difficult to use.

We have considered different ways of using the frequent set to obtain more useful information. One interesting task is to find similarities between attributes on the basis of context information. The basic idea is that two 0-1 attributes are similar, if they occur in similar contexts. For example, in market basket data Coke and Pepsi can be considered to be similar, if the buying behavior of Coke buyers is similar to the buying behavior of Pepsi buyers. An iterative algorithm based on this idea is given in [7]. More detailed probabilistic modeling of market basked data is considered in [8].

Random projection in dimensionality reduction

In many applications of data mining, the high dimensionality of the data restricts the choice of data processing methods. A statistically optimal way of dimensionality reduction is to project the data onto a lower-dimensional orthogonal subspace that captures as much of the variation of the data as possible. The best (in mean-square sense) and most widely used way to do this is principal component analysis (PCA); unfortunately it is quite expensive to compute for high-dimensional data sets, as it requires the eigenvalue decomposition of the data covariance matrix.

A computationally simple method of dimensionality reduction that does not introduce a significant distortion in the data set would thus be desirable. Random projection (RP) has recently emerged as a promising alternative for this task. In random projection, the original d-dimensional data matrix X is projected onto a k-dimensional ($k \ll d$) subspace through the origin, using a random $k \times d$ matrix R whose columns have unit lengths. Using matrix notation where $X_{d\times N}$ is the original set of N d-dimensional observations,

$$X^{RP}_{k\times N} = R_{k\times d}X_{d\times N} \tag{19.1}$$

is the projection of the data onto a lower k-dimensional subspace. The key idea of random mapping arises from the Johnson-Lindenstrauss lemma [9]: if points in a vector space are projected onto a randomly selected subspace of suitably high dimension, then the distances between the points are approximately preserved.

While this method has attracted lots of interest, empirical results are sparse. For this reason, we have presented experimental results on using random projection as a dimensionality reduction tool in the context of both noisy and noiseless images, and information retrieval in text documents [10]. As an example, we present here results on monochrome images of natural scenes. The distortion caused by dimensionality reduction is measured by comparing the Euclidean distance between two dimensionality reduced data vectors to their Euclidean distance in the original high-dimensional space. Figure 19.1 shows the mean squared error in the distance between members of a pair of data vectors, averaged over 100 pairs. The results of random projection with a Gaussian distributed random matrix (RP), random projection with a sparse random matrix [11] (SRP), principal component analysis (PCA) and discrete cosine transform (DCT) are shown, together with their 95 per cent confidence intervals. Figure 19.2 shows the number of Matlab's floating point operations needed when using the abovementioned methods. It can be seen that PCA is significantly more burdensome than RP or DCT.

To conclude, we have shown that projecting the data onto a random lower-dimensional subspace yields results comparable to conventional dimensionality reduction methods such as principal component analysis: the similarity of data vectors is preserved well under random projection. We have also shown experimentally that using a sparse random matrix [11] gives additional computational savings in random projection.

References

[1] D. Hand, H. Mannila and P. Smyth. *Principles of Data Mining.* MIT Press 2001. ISBN 0-262-98290-X.

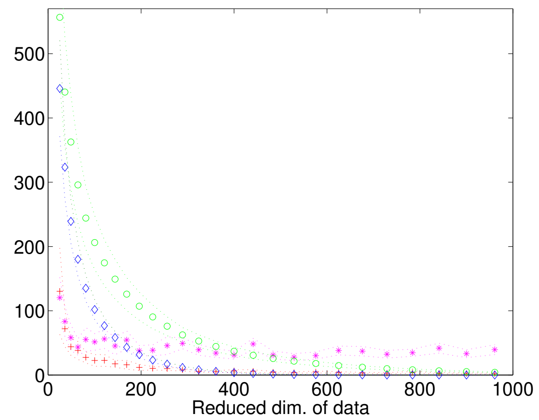

Figure 19.1: MSE of Euclidean distances, and 95 % confidence intervals over 100 pairs of image vectors. RP+ sparse RP$*$ PCA$\diamond$ DCT$\circ$

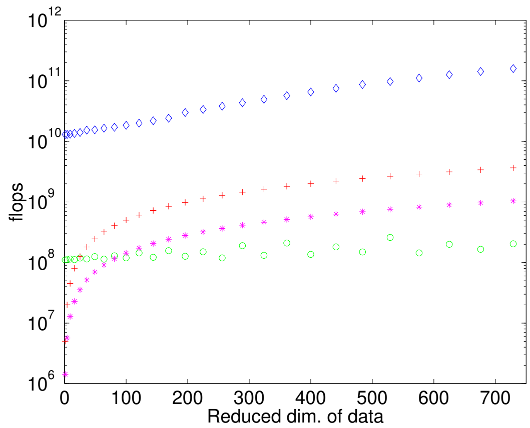

Figure 19.2: Number of Matlab's floating point operations needed when reducing the dimensionality of image data using RP+ sparse RP$*$ PCA$\diamond$ DCT$\circ$

[2] H. Mannila and C. Meek. Global partial orders from sequential data. *Sixth Annual Conference on Knowledge Discovery and Data Mining (KDD-2000)*, pp. 161–168.

[3] H. Mannila and J.K. Seppänen. Recognizing similar situations from event sequences. *First SIAM Conference on Data Mining*, 2001. `http://www.siam.org/meetings/sdm01/pdf/sdm01_03.pdf`

[4] H. Mannila and D. Rusakov. Decomposing event sequences into independent components. *First SIAM Conference on Data Mining*, 2001. `http://www.siam.org/meetings/sdm01/pdf/sdm01_02.pdf`

[5] H. Mannila and M. Salmenkivi. Finding simple intensity descriptions from event sequence data. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001)*, F. Provost and R. Srikant (eds.), pp. 341–346.

[6] J. Himberg, K. Korpiaho, H. Mannila, J. Tikanmäki and H. Toivonen. Time-series segmentation for context recognition in mobile devices. *IEEE Conference on Data Mining* 2001, pp. 203–207.

[7] G. Das and H. Mannila. Context-based similarity methods for categorical attributes. *Principles of Data Mining and Knowledge Discovery, 4th European Conference (PKDD 2000)*, D.A. Zighed et al. (eds.), pp. 201–211.

[8] I. Cadez, P. Smyth and H. Mannila. Probabilistic Modeling of Transaction Data with Applications to Profiling, Visualization, and Prediction. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001)*, F. Provost and R. Srikant (eds.), pp. 37–46.

[9] W.B. Johnson and J. Lindenstrauss. Extensions of Lipshitz mapping into Hilbert space. In *Conference in modern analysis and probability*, volume 26 of *Contemporary Mathematics*, pages 189–206. Amer. Math. Soc., 1984.

[10] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proc. 7th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD-2001)*, pages 245–250, 2001.

[11] D. Achlioptas. Database-friendly random projections. In *Proc. ACM Symp. on the Principles of Database Systems*, pages 274–281, 2001.

[12] C.K. Leung, R. Ng, and H. Mannila. Segmentation for frequency counting. *ICDE 2002*, to appear.

[13] H. Mannila. Theoretical frameworks for data mining. *SIGKDD Explorations* 1, 2 (January 2000), pp. 30–32

[14] J. Mäntyjärvi, J. Himberg, P. Korpipää, and H. Mannila. Extracting the context of a mobile device user. To appear.

19.2 Applications in bioinformatics

Gene Expression

The research of the Pattern Discovery group in genetic expression data has been concentrated in two main directions. First, we have examined publically available data on expression in baker's yeast in order to understand the fundaments of the field and to develop new techniques; second, we have applied proven techniques to data on expression in humans.

All biological organisms contain genetic material in the form of DNA molecules. These molecules are copied in the cell division process, and thus all cells of a single organism contain copies of the same DNA. The differences in cell function are caused by differential *expression* of the genes. Genes are expressed in a process where DNA is copied into RNA, which in turn guides the manufacturing of proteins. The natural control mechanisms of expression are largely unknown: some low-level building blocks have been determined, but uncovering regulatory networks of genes remains a big challenge.

DNA microarrays provide a technical method for measuring genetic expression. Single-stranded stretches of DNA are spotted onto a glass slide; each spot contains some DNA from a single gene, and is usually repeated at a few different locations on the slide. In the actual experiment, messenger RNA (mRNA) is obtained from the two tissue samples under consideration, complementary DNA (cDNA) is produced by a biochemical reverse transcription mechanism, and each of the two samples is labeled with a different fluorescent dye. These labeled cDNA samples are then allowed to hybridize on the slide, and the amounts of the two fluorescent dyes are measured with a confocal laser scanner. If a gene has differential expression in the two tissues, the original samples contain different amounts of mRNA, and thus different amounts of labeled cDNA ends up at the spot corresponding to the gene. This will cause a difference in the intensities measured by the scanner.

DNA microarrays can measure the expression levels of thousands of genes in parallel, whereas older methods require a separate experiment for each gene. They thus promise to help in the task of reverse engineering genetic regulation networks.

In medical genetics, the two compared tissues typically represent healthy and sick patients, for example normal lung cells and cancer cells. Also, e.g., the effects of a treatment can be monitored by using untreated tissue as reference and samples taken at different time points after the treatment; this yields time series data.

Experiments on Yeast Data

We examined several publically available datasets on expression in baker's yeast (*Saccharomyces cerevisiae*). Our statistical analysis indicates a correlation between genes located in the same chromosome that is only partially explained by known regulation mechanisms. These mechanisms function at a small spatial range, and indeed genes that are located close to each other are more tightly coregulated; but also genes far away from each other show a small but significant correlation. This could be explained by a hitherto unknown regulation mechanism or possibly by some chromosomally systematic bias in the microarray experiments. [1]

Human Data: Cancer and Asthma

We have collaborated with two research groups: one in the Centre of Biotechnology in Turku, led by Riitta Lahesmaa, and another in the Department of Medical Genetics at the University of Helsinki, led by Sakari Knuutila. The collaboration with Centre of

Biotechnology researchers has centered on validation of the Finnish microarray design intended for research into the genetic causes of asthma. We have given feedback on the quality of data obtained by researchers. This feedback has resulted in parts of some experiments being repeated on another scanner.

Together with the group of Sakari Knuutila, we have analyzed genetic expression data from patients suffering from mantle cell lymphoma [4], childhood leukemia [2], lung adenocarcinoma [3], and lung small cell carcinoma. The methods we have used have this far consisted of fairly standard statistics combined with computer-intensive permutation tests and ROC curve analyses. After these basic studies, we aim to continue with more ambitious clustering and other data analysis methods.

Gene Mapping

One of the large challenges in current medical genetics is locating genes predisposing to common multifactorial diseases. While there have been many successes in finding the genes responsible for various rare diseases, finding the genetic causes of common diseases such as asthma and diabetes poses still large problems.

Work done in colllaboration with the Finnish Genome Center and the Department of Computer Science of the University of Helsinki has developed methods for analyzing multifactorial diseases, for helping in study design, and in finding candidate loci for genes in multifactorial diseases.

One of the key problems in finding the genes predisposing to common traits is that the nature of genetic models is not well understood. A long-term goal in the group is to develop methods for deriving the potential genetic models from epidemiological data; the tools in this study are mainly Markov Chain Monte Carlo techniques. One recent result related to this study is a proof that any (single or multilocus) genetic model the sibling risk is at least as high as the offspring risk [5]. While the result is simple in the case of a single locus, the proof is highly nontrivial for the general case.

The design of empirical studies for finding genes contributing to certain traits is highly difficult. The population simulator developed in the same collaboration makes it possible to estimate whether a certain study design is appropriate for certain traits, and to estimate the significance of findings of studies; the tool has been used in a variety of studies [6, 7]. The haplotype pattern mining (HPM) method uses pattern discovery techniques to find the probable loci of predisposing genes. The performance of the method is extremely good [8], and it can also be generalized in a variety of ways [9].

References

[1] Heikki Mannila, Anne Patrikainen, Jouni K. Seppänen, and Juha Kere. Long-range control of expression in yeast. Accepted for publication in *Bioinformatics*.

[2] Tarja Niini, Jaakko Hollmén, Marcelo L Larramendy, Yan Aalto, Balint Nagy, Kim Vettenranta, and Sakari Knuutila. Gene Expression Profiling in Childhood Acute Lymphoblastic Leukaemia by cDNA array. Poster presented in *Beatson International Cancer Conference*, Genomic Regulation and Cancer, July 2001, Glasgow, Scotland.

[3] Harriet Wikman, Eeva Kettunen, Jouni K. Seppänen, Jaakko Hollmén, Antti Karjalainen, Sisko Anttila, and Sakari Knuutila. Gene Expression Profiling of Adenocarcinoma of the Lung. *Abstracts of the Asian-Pacific Conference on Tumor Biology*, September 2001, Peking, China.

[4] Ying Zhu, Jaakko Hollmén, Riikka Oinonen, Kaarle Franssila, Yan Aalto, Erkki Elonen, Heikki Mannila, Juha Kere, and Sakari Knuutila. Proffered paper presented in European Cancer Conference ECCO 11, October 2001, Lisbon. *European Journal of Cancer*, Vol. 37 Supplement 6, p. S41.

[5] M. Koivisto and H. Mannila. Offspring risk and sibling risk for multilocus traits. *Human Heredity* 51 (2001), 209–216.

[6] P. Kauppi, T. Laitinen, V. Ollikainen, H. Mannila, L.A. Laitinen, and J. Kere. The ILR9 region contribution in asthma is supported by genetic association in an isolated population. *European Journal of Human Genetics* 8, 788–792 (2000).

[7] T. Laitinen, V. Ollikainen, C. Lazaro, P. Kauppi, R de Cid, J.M. Anto, X. Estivill, H. Lokki, H. Mannila, L.A. Laitinen and J. Kere. Association study of the chromosomal region containing the FCER2 gene suggests it has a regulatory role in atopic disorders. *American Journal on Respiratory and Critical Care Medicine* 161, 700–706, 2000.

[8] H. Toivonen, P. Onkamo, K. Vasko, V. Ollikainen, P. Sevon, H. Mannila, M. Herr, and J. Kere. Data mining applied to linkage disequilibrium mapping. *American Journal of Human Genetics* 67(1): 133–145, July 2000.

[9] P. Sevon, V. Ollikainen, P. Onkamo, H. Toivonen, H. Mannila and J. Kere. Mining associations between genetic markers, phenotypes and covariates. *Genetic Epidemiology*, 21(Suppl 1): S588–S593, 2001.