II From Data to Knowledge Research Unit Research Projects under the CIS Laboratory

Chapter 17

From Data to Knowledge Research Unit

Heikki Mannila, Jaakko Hollmén, Ella Bingham, Johan Himberg, Mikko Koivisto, Anne Patrikainen, Salla Ruosaari, Jouni K. Seppänen, Mikko Katajamaa, Heli Juntunen, Nikolaj Tatti, Antti Rasinen, Kalle Korpiaho, Jaripekka Juhala, Antti Savolainen, Mikko Korpela, Janne Toivonen

17.1 Data mining

The work 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. One of the basic questions is how much does the collection of frequent patterns tell us about the underlying distribution. We have analyzed the use of maximum entropy approaches to inferring distributions from frequent pattern collections and obtained quite good empirical results [6]. Another major question is finding structure in large collection of 0-1 data: the results include a simple model of topics in 0-1 data, and simple algorithms for finding the topic structure [5]. In industrial cooperation projects we have recently developed simple and efficient algorithms for on-line clustering.

The combination of probabilistic and algorithmic techniques is also visible in several new themes. One major new theme in the work is in finding good segmentations for sequences. The (k,h)-segmentation problem and algorithms [2] show how one can locate recurrent sources from sequences; the approach applies to basically any probabilistic model for the generation of points in the sequences. We have also looked at the question of finding fragments of total orders from unordered data [1], which seems to be a fruitful approach. We are also investigating different approaches to subspace clustering.

On pure pattern discovery area, topics include approximation of frequent set collections and pattern discovery algorithms [3].

The work on combining local and global analysis in data mining will continue. Potential new themes include spectral clustering, interplay of probabilistic clustering and frequent sets [4], and word discovery from sequences. The work has lots of connections to applications, e.g., in paleontology and genomics.

References

[1] A. Gionis, T. Kujala and H. Mannila: Fragments of order. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* P. Domingos et al. (eds.), pages 129–136, 2003.

[2] A. Gionis and H. Mannila: Finding recurrent sources in sequences. *The Seventh Annual International Conference on Research in Computational Molecular Biology*, W. Miller, M. Vingron, S. Istrail, P. Pevzner, M. Waterman (eds.), pages 123–130, ACM, 2003.

[3] D. Gunopulos, R. Khardon, H. Mannila, S. Saluja, H. Toivonen, and R.S. Sarma. Discovering all most specific sentences. *ACM Transactions on Database Systems* (28)2:140–174, 2003.

[4] Jaakko Hollmén, Jouni K. Seppänen, and Heikki Mannila. Mixture models and frequent sets: combining global and local methods for 0-1 data. In Daniel Barbará and Chandrika Kamath, editors, *Proceedings of the Third SIAM International Conference on Data Mining*, pages 289–293. Society of Industrial and Applied Mathematics, 2003.

[5] Jouni K. Seppänen, Ella Bingham, and Heikki Mannila. A simple algorithm for topic identification in 0-1 data. In Nada Lavrač, Dragan Gamberger, Ljupčo Todorovski, and Hendrik Blockeel, editors, *Knowledge Discovery in Databases: PKDD 2003. 7th European Conference on Principles and Practice of Knowledge Discovery*

in Databases. Cavtat-Dubrovnik, Croatia, September 2003, Proceedings, number 2838 in Lecture Notes in Artificial Intelligence, pages 423–434. Springer, 2003.

[6] D. Pavlov, H. Mannila, and P. Smyth. Beyond independence: probabilistic methods for query approximation on binary transaction data. *IEEE Transactions on Data and Knowledge Engineering* (15)6:1409–1421, 2003

17.2 Latent topics in 0-1 data

Large collections of 0–1 data occur in many applications, such as information retrieval, web browsing, telecommunications, and market basket analysis. While the dimensionality of such data sets can be large, the variables (or attributes) are seldom completely independent. Rather, it is natural to assume that the attributes are organized into (possibly overlapping) *topics*, i.e., collections of variables whose occurrences are somehow connected to each other. For example, in document data the topics correspond to topics of the document: e.g., phrases "data mining", "decision trees" and "association rules" probably are included in one topic, which might be called the "data mining" topic. In supermarket market basket data, the topics could correspond to classes of products such as soft drinks, vegetables, etc. In discretized gene expression data topics could correspond to groups of genes that are expressed in similar conditions or tissues.

Finding topics from data is by no means easy: the topics can be overlapping; a particular topic may be active only for a subset of documents; all attributes in a topic might not be present in the same observation. In the papers [1] and [2] we describe methods to estimate these hidden topics in 0-1 data. We specify several data models and give algorithms for finding the topics. An example of our topic model is given in Figure 17.1. The observed data are generated by interactions between independent latent topics: Each topic has a probability s_j of being active in an observation vector. The topics j then generate occurrences of variables $A, B, C, \ldots$ according to some topic-variable probabilities that are listed in matrix $\mathbf{A}$.

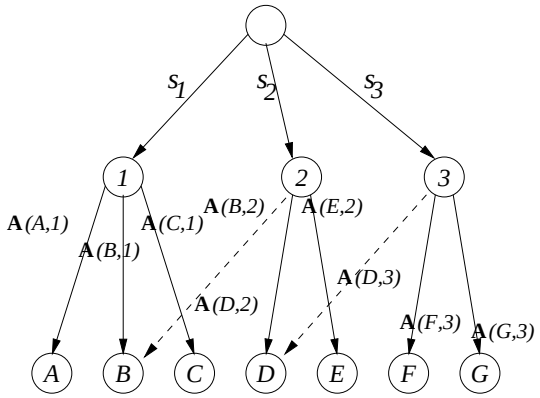

Figure 17.1: An example topic model. Topics 1, 2 and 3 are generated independently of each other with probabilities s_1, s_2 and s_3. The topics then generate observed variables with probabilities $\mathbf{A}(i, j)$. The dashed arrows indicate that a variable may be generated by several topics.

References

[1] Ella Bingham, Heikki Mannila, and Jouni K. Seppänen. Topics in 0-1 data. In David Hand, Daniel Keim, and Raymond Ng, editors, *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 450–455, Edmonton, Alberta, Canada, July 2002.

[2] Jouni K. Seppänen, Ella Bingham, and Heikki Mannila. A simple algorithm for topic identification in 0-1 data. In Nada Lavrač, Dragan Gamberger, Ljupčo Todorovski, and Hendrik Blockeel, editors, *Knowledge Discovery in Databases: PKDD*

2003. 7th European Conference on Principles and Practice of Knowledge Discovery in Databases. Cavtat-Dubrovnik, Croatia, September 2003, Proceedings, number 2838 in Lecture Notes in Artificial Intelligence, pages 423–434. Springer, 2003.

17.3 Applications in bioinformatics

Gene expression data analysis

Thousands of simultaneous gene expression measurements can be obtained with the microarray platform. As a result of an experiment, the analysts are faced with an abundance of data, usually a $N \times p$ matrix of continuous gene expression measurements, where N denotes the number of measured genes (usually in the thousands) and p the number of samples or subjects (usually a handful). In close collaboration with specialists in relevant fields in biology and medicine, we have analyzed this type of gene expression data in various cancer investigations [1,3,6,7], where the patient material is well characterized and other information is also available.

The analysis of the resulting gene expression data matrix can proceed in many alternative ways. We have used signal decomposition methods, for instance, principal component analysis [3,6,7] and non-negative matrix factorization [5] to yield meaningful components from the data. For instance, projections of data on the first principal component have been used as a score for collective difference in expression between the samples and the reference. In order to avoid stating random findings as true, we have extensively used the permutation testing in validating the results with the data set at hand [1,2,3,6,7]. Findings from the screening type of studies should be externally confirmed [1,3,6,7].

We have also examined publicly available gene expression data from baker's yeast [2]. 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 co-regulated; but also genes far away from each other show a small but significant correlation. By analyzing gene expression data in combination with other sources of data, one can make improved inferences.

Currently, the work continues with method development in the probabilistic framework to combine several sources of data, and to draw improved inferences based on the joint data set. The immediate application area is found in our collaborative cancer research: we are working on a project aiming at finding tumor markers of work-related lung cancers. Existing measurements include gene expression data from the microarray platform, copy number alteration measurements along the chromosome, characterization of the patient material, and gene annotation databases.

Quality control of microarray data

One strand of our work has investigated quality control of data originating from the microarray measurement platform, based on image analysis of scanned images of hybridized microarrays [4]. The goal is to be able to automatically classify spots into good and faulty spots, so that no erroneous spot would enter the subsequent analysis, therefore possibly causing bias in results. We extract features from the spot image describing shape, regularity and uniformity, and train a Naive Bayes classifier on the extracted features using a pre-labeled database of spot images. Furthermore, we describe a non-symmetric cost model in the cost-sensitive classification setting. Out of the three repetitions of the same measurement, we should allow as many good measurements as possible to enter to subsequent analysis, but to prevent an erroneous spot to be taken into account in the analysis phase. The results are assessed with Receiver Operating Characteristic (ROC) curves in the classification setting and expected costs in the cost-sensitive classification setting.

We plan to investigate the extension of these techniques to a three-color microarray platform, where a third channel is used to effectively bind to all probes of the array.

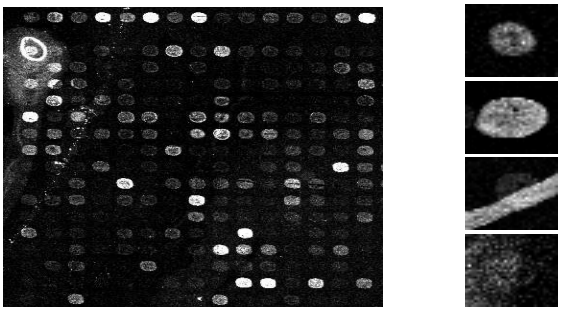

Figure 17.2: Example of an image of microarray containing numerous faulty spots is shown left. Four examples on the images on the right side demonstrate possible problems, for instance, spots of varying sizes in the two upper images and scratches and noise in the two lower images, respectively.

References

[1] Eeva Kettunen, Sisko Anttila, Jouni K. Seppänen, Antti Karjalainen, Henrik Edgren, Irmeli Lindström, Reijo Salovaara, Anna-Maria Nissén, Jarmo Salo, Karin Mattson, Jaakko Hollmén, Sakari Knuutila, and Harriet Wikman. Differentially expressed genes in non-small cell lung cancer (NSCLC) expression profiling of cancer-related genes in squamous cell lung cancer. *Cancer Genetics and Cytogenetics*, In press.

[2] Heikki Mannila, Anne Patrikainen, Jouni K. Seppänen, and Juha Kere. Long-range control of expression in yeast. *Bioinformatics*, 18(3):482–483, 2002.

[3] Tarja Niini, Kim Vettenranta, Jaakko Hollmén, Marcelo L. Larramendy, Yan Aalto, Harriet Wikman, Bálint Nagy, Jouni K. Seppänen, Anna Ferrer Salvador, Heikki Mannila, Ulla M. Saarinen-Pihkala, and Sakari Knuutila. Expression of myeloid-specific genes in childhood acute lymphoblastic leukemia — a cDNA array study. *Leukemia*, 16(11):2213–2221, 2002. Nature Publishing Group.

[4] Salla Ruosaari and Jaakko Hollmén. Image analysis for detecting faulty spots from microarray images. In Steffen Lange, Ken Satoh, and Carl H. Smith, editors, *Proceedings of the 5th International Conference on Discovery Science (DS 2002)*, volume 2534 of *Lecture Notes in Computer Science*, pages 259–266. Springer-Verlag, 2002.

[5] Jouni K. Seppänen, Jaakko Hollmén, Ella Bingham, and Heikki Mannila. Nonnegative Matrix Factorization on Gene Expression Data. *Bioinformatics 2002*, Bergen, Norway, April 4-7, 2002. poster 49.

[6] Harriet Wikman, Eeva Kettunen, Jouni K. Seppänen, Antti Karjalainen, Jaakko Hollmén, Sisko Anttila, and Sakari Knuutila. Identification of differentially expressed genes in pulmonary adenocarcinoma by using a cDNA array. *Oncogene*, 21(37):5804–5813, 2002. Nature Publishing Group.

[7] Ying Zhu, Jaakko Hollmén, Riikka Räty, Yan Aalto, Balint Nagy, Erkki Elonen, Juha Kere, Heikki Mannila, Kaarle Franssila, and Sakari Knuutila. Investigatory and analytical approaches to differential gene expression profiling in mantle cell lymphoma. *British Journal of Haematology*, 119(4):905–915, 2002.