Publications of the From Data to Knowledge Research Unit

[1] E. Bingham, A. Gionis, N. Haiminen, H. Hiisilä, H. Mannila, and E. Terzi. Segmentation and dimensionality reduction. In *2006 SIAM Conference on Data Mining 2006*, pages 372–383, 2006.

[2] A. Dasgupta, G. Das, and H. Mannila. A random walk approach to sampling hidden databases. In *Proc. of the 2007 ACM SIGMOD International Conference on Management of Data (SIGMOD 2007)*, pages 629–640, 2007.

[3] G. C. Garriga, H. Heikinheimo, and J. K. Seppänen. Cross-mining binary and numerical attributes. In *Proc. of IEEE International Conference on Data Mining (ICDM)*, pages 481–486, 2007.

[4] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(3):14, 2007.

[5] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. In *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2006*, pages 167–176, 2006.

[6] A. Gionis, H. Mannila, K. Puolamäki, and A. Ukkonen. Algorithms for discovering bucket orders from data. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 561–566, 2006.

[7] A. Gionis, H. Mannila, and P. Tsaparas. Clustering aggregation (long version). *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[8] R. Gupta, S. Ruosaari, S. Kulathinal, J. Hollmén, and P. Auvinen. Microarray image segmentation using additional dye - an experimental study. *Molecular and Cellular Probes*, (5-6):321–328, 2007.

[9] R. Gwadera, A. Gionis, and H. Mannila. Optimal segmentation using tree models. In *2006 IEEE International Conference on Data Mining 2006*, pages 244–253, 2006.

[10] R. Gwadera, J. Toivola, and J. Hollmén. Segmenting multi-attribute sequences using dynamic Bayesian networks. In *Proceedings of The Seventh IEEE International Conference on Data Mining - Workshops (ICDM Workshops 2007)*, pages 465–470. IEEE Computer Society, 2007.

[11] N. Haiminen and H. Mannila. Discovering isochores by least-squares optimal segmentation. *Gene*, 394(1-2):53–60, 2007.

[12] N. Haiminen, H. Mannila, and E. Terzi. Comparing segmentations by applying randomization techniques. *BMC Bioinformatics*, 8(171 (23 May 2007)), 2007.

[13] D. R. Hardoon, J. Shawe-Taylor, A. Ajanki, K. Puolamäki, and S. Kaski. Information retrieval by inferring implicit queries from eye movements. In M. Meila and X. Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, 2007.

[14] H. Heikinheimo, M. Fortelius, J. Eronen, and H. Mannila. Biogeography of european land mammals shows environmentally distinct and spatially coherent clusters. *Journal of Biogeography*, 34(6):1053–1064, 2007.

[15] H. Heikinheimo, E. Hinkkanen, H. Mannila, T. Mielikäinen, and J. K. Seppänen. Finding low-entropy sets and trees from binary data. In *Proc. of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 350–359, 2007.

[16] H. Heikinheimo, H. Mannila, and J. K. Seppänen. Finding trees from unordered 0-1 data. In J. Fürnkranz, T. Scheffer, and M. Spiliopoulou, editors, *Knowledge Discovery in Databases: PKDD 2006*, pages 175–186. Springer, 2006.

[17] A. Hinneburg, H. Mannila, S. Kaislaniemi, T. Nevalainen, and H. Raumolin-Brunberg. How to handle small samples: Bootstrap and Bayesian methods in the analysis of linguistic change. *Literary and Linguistic Computing*, 22(2):137–150, 2007.

[18] J. Hollmén. Model selection and estimation via subjective user preferences. In V. Corruble, M. Takeda, and E. Suzuki, editors, *Proceedings of The Tenth International Conference on Discovery Science (DS 2007)*, pages 259–263, Sendai, 2007. Springer-Verlag.

[19] J. Hollmén and J. Tikka. Compact and understandable descriptions of mixture of Bernoulli distributions. In M. Berthold, J. Shawe-Taylor, and N. Lavrac, editors, *Proceedings of the 7th International Symposium on Intelligent Data Analysis (IDA 2007)*, pages 1–12, Ljubljana, 2007. Springer-Verlag.

[20] S. Hyvönen, A. Gionis, and H. Mannila. Recurrent predictive models for sequence segmentation. In *Advances in Intelligent Data Analysis VII (IDA 2007) 2007*, pages 195–206, 2007.

[21] H. Keski-Säntti, T. Atula, J. Tikka, J. Hollmén, A. A. Mäkitie, and I. Leivo. Predictive value of histopathologic parameters in early squamous cell carcinoma of oral tongue. *Oral Oncology*, 43(10):1007–1013, 2007.

[22] M. Korpela and J. Hollmén. Extending an algorithm for clustering gene expression time series. In J. Rousu, S. Kaski, and E. Ukkonen, editors, *Probabilistic Modeling and Machine Learning in Structural and Systems Biology 2006*, pages 120–124, Helsinki, 2006.

[23] N. Landwehr, T. Mielikäinen, L. Eronen, H. Toivonen, and H. Mannila. Constrained hidden markov models for population-based haplotyping. *BMC Bioinformatics*, 8 (Suppl 2):1–9, 2007.

[24] N. Landwehr, T. Mielikäinen, L. Eronen, H. Toivonen, and H. Mannila. Constrained hidden markov models for population-based haplotyping. In *Probabilistic Modeling and Machine Learning in Structural and Systems Biology (PMSB 2006) 2006*, 2006.

[25] P. Lindholm, P. Nymark, H. Wikman, K. Salmenkivi, A. Nicholson, M. V. Korpela, S. Kaski, S. Ruosaari, J. Hollmen, E. Vanhala, A. Karjalainen, S. Anttila, V. Kinnula, and S. Knuutila. Asbestos-associated malignancies in the lung and pleura show distinct genetic aberrations. *Lung cancer*, 54:15, 2006.

[26] S. Luyssaert, I. Janssens, M. Sulkava, D. Papale, A. Dolman, M. Reichstein, T. Suni, J. Hollmén, T. Vesala, D. Lousteau, B. Law, and E. Moors. Photosynthesis drives interannual variability in net carbon-exchange of pine forests at different latitudes. In *Proceedings of the Open Science Conference on the GHG Cycle in the Northern Hemisphere*, pages 86–87, Jena, Germany, 2006. CarboEurope, NitroEurope, CarboOcean, and Global Carbon Project.

[27] S. Luyssaert, I. A. Janssens, M. Sulkava, D. Papale, A. J. Dolman, M. Reichstein, J. Hollmén, J. G. Martin, T. Suni, T. Vesala, D. Lousteau, B. E. Law, and E. J. Moors. Photosynthesis drives anomalies in net carbon-exchange of pine forests at different latitudes. *Global Change Biology*, 13(10):2110–2127, 2007.

[28] S. Luyssaert, M. Sulkava, H. Raitio, J. Hollmén, and P. Merilä. Is n and s deposition altering the mineral nutrient composition of norway spruce and scots pine needles in finland? In J. Eichhorn, editor, *Proceedings of Symposium: Forests in a Changing Environment - Results of 20 years ICP Forests Monitoring*, pages 80–81, Göttingen, Germany, 2006. ICP Forests, European Commission, Nordwestdeutsche Versuchsanstalt.

[29] H. Mannila. The role of information technology for systems biology. in systems biology: A grand challenge for europe, esf 2007, p. 21-23., 2007.

[30] H. Mannila and E. Terzi. Nestedness and segmented nestedness. In *Proc. of 13th ACM SIGKDD International Conference on Knowledge*, pages 480–489, 2007.

[31] P. Miettinen, T. Mielikäinen, A. Gionis, G. Das, and H. Mannila. The discrete basis problem. In *10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD) 2006 2006*, 2006.

[32] S. Myllykangas, J. Himberg, T. Böhling, B. Nagy, J. Hollmén, and S. Knuutila. Dna copy number amplification profiling of human neoplasms. *Oncogene*, 25(55):7324–7332, 2006.

[33] P. Nymark, P. M. Lindholm, M. V. Korpela, L. Lahti, S. Ruosaari, S. Kaski, J. Hollmen, S. Anttila, V. L. Kinnula, and S. Knuutila. Gene expression profiles in asbestos-exposed epithelial and mesothelial lung cell lines. *BMC Genomics*, 8:1–14, 2007.

[34] P. Nymark, H. Wikman, S. Ruosaari, J. Hollmén, E. Vanhala, A. Karjalainen, S. Anttila, and S. Knuutila. Identification of specific gene copy number changes in asbestos-related lung cancer. *Cancer Research*, 66(11):5737–5743, 2006.

[35] K. Puolamäki, M. Fortelius, and H. Mannila. Seriation in paleontological data using markov chain monte carlo methods. *PLoS Computational Biology*, 2(2):26, 2006.

[36] K. Puolamäki, M. Fortelius, and H. Mannila. Seriation in paleontological data using markov chain monte carlo methods, 2006.

[37] K. Puolamäki, S. Hanhijärvi, and G. C. Garriga. An approximation ratio for biclustering. Technical Report Report E13, Espoo, Finland, 2007.

[38] K. Puolamäki and S. Kaski. *Proceedings of the NIPS 2005 Workshop on Machine Learning for Implicit Feedback and User Modeling*. Espoo, 2006.

[39] K. Puolamäki, J. Salojärvi, E. Savia, and S. Kaski. Discriminative mcmc. Technical Report Report E1, Helsinki University of Technology, Espoo, Finland, 2006.

[40] A. Rasinen, J. Hollmén, and H. Mannila. Analysis of linux evolution using aligned source code segments. In N. Lavrac, L. Todorovski, and K. Jantke, editors, *Proceedings of the Ninth International Conference on Discovery Science*, pages 209–218. Springer-Verlag, 2006.

[41] E. Salmela, O. Taskinen, J. K. Seppänen, P. Sistonen, M. J. Daly, P. Lahermo, M.-L. Savontaus, and J. Kere. Subpopulation difference scanning: a strategy for exclusion mapping of susceptibility genes. *Journal of medical genetics*, 43:590–597, 2006.

[42] M. Sulkava, S. Luyssaert, P. Rautio, I. A. Janssens, and J. Hollmén. Modeling the effects of varying data quality on trend detection in environmental monitoring. *Ecological Informatics*, 2(2):167–176, 2007.

[43] M. Sulkava, H. Mäkinen, P. Nöjd, and J. Hollmén. CUSUM charts for detecting onset and cessation of xylem formation based on automated dendrometer data. In I. Horová and J. Hrebícek, editors, *TIES 2007 - 18th annual meeting of the International Environmetrics Society*, page 111, Mikulov, 2007. Masaryk University.

[44] M. Sulkava, J. Tikka, and J. Hollmén. Sparse regression for analyzing the development of foliar nutrient concentrations in coniferous trees. *Ecological Modelling*, 191:118–130, 2006.

[45] N. Tatti. Computational complexity of queries based on itemsets. *Information Processing Letters*, pages 183–187, 2006.

[46] N. Tatti. Safe projections of binary data sets. *Acta Informatica*, 42(8-9):617–638, 2006.

[47] N. Tatti. Distances between data sets based on summary statistics. *Journal of Machine Learning Research*, 8:131–154, 2007.

[48] N. Tatti. Maximum entropy based significance of itemsets. In *Proc. of Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 312–321, 2007.

[49] N. Tatti, T. Mielikäinen, A. Gionis, and H. Mannila. What is the dimension of your binary data? In *2006 IEEE International Conference on Data Mining 2006*, pages 603–612, 2006.

[50] J. Tikka and J. Hollmén. Long-term prediction of time series using a parsimonious set of inputs and LS-SVM. In A. Lendasse, editor, *Proceedings of the First European Symposium on Time Series Prediction (ESTSP 2007)*, pages 87–96, Espoo, 2007.

[51] J. Tikka and J. Hollmén. A sequential input selection algorithm for long-term prediction of time series. *Neurocomputing*, 2007.

[52] J. Tikka, J. Hollmén, and S. Myllykangas. Mixture modeling of DNA copy number amplification patterns in cancer. In F. Sandoval, A. Prieto, J. Cabestany, and M. Grana, editors, *Proceedings of the 9th International Work-Conference on Artificial Neural Networks (IWANN 2007)*, pages 972–979, San Sebastián, 2007.

[53] J. Tikka, A. Lendasse, and J. Hollmén. Analysis of fast input selection: Application in time series prediction. In *International Conference on Artificial Neural Networks (ICANN) 2006*, pages 161–170, 2006.

[54] A. Ukkonen. Visualizing sets of partial rankings. In *Proc. of Advances in Intelligent Data Analysis VII, 7th International Symposium on Intelligent Data Analysis (IDA 2007)*, pages 240–251, 2007.

[55] A. Ukkonen and H. Mannila. Finding outlying items in sets of partial rankings. In *Proc. of PKDD 2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 265–276, 2007.

[56] H. Wikman, S. Ruosaari, P. Nymark, V. K. Sarhadi, J. Saharinen, E. Vanhala, A. Karjalainen, J. Hollmén, S. Knuutila, and S. Anttila. Gene expression and copy number profiling suggests the importance of allelic imbalance in 19p in asbestos-associated lung cancer. *Oncogene*, 26(32):4730–4737, 2007.