On the robustness of semi-supervised hierarchical graph clustering in functional genomics

Johann M. Kraus [⋆], Günther Palm [⋆], Hans A. Kestler [†,⋆]

[⋆]Institute of Neural Information Processing, University of Ulm, 89069 Ulm
[†]Department of Internal Medicine I, University Hospital Ulm, 89081 Ulm

1 Introduction

Clustering is an important technique in the explorative data analysis and is applied in many different disciplines such as text mining, marketing, psychology and biology. The aim of cluster analysis is to identify either a grouping into a specified number of clusters or a hierarchy of nested partitions. In contrast to supervised machine learning techniques cluster analysis does not need a teacher signal. Therefore clustering is often used when no knowledge about the structure of the data is present. In some tasks, however, limited background knowledge may be available. This might be a set of labels for a small amount of data or known relations between some data items. Wagstaff and Cardie [1] suggest to model the relationship between pairs of data items as must-link and cannot-link constraints. Must-links indicate that two objects must be in the same cluster and on the contrary cannot-links denote that two objects should not be grouped together.

The focus of our research interest lies on the functional grouping of genes or the preclassification of gene profiles. Functional genomics as part of molecular biology aims to provide a link between genomic information and biological functions as for example determining a relation between gene expression patterns and tumor status. Microarray data mining is an important issue in studying biological processes. As hierarchical clustering algorithms can predict a basic branching structure they are often used in this context [2]. Limitations in the reproducibility of clustering results after small modifications of the data set motivated us to include background knowledge into the hierarchical clustering process. Figure 1 shows a dendrogram resulting from clustering microarray data where removing one data item leads to a markedly change of the cluster result. In previous work [3] we showed that clustering became more robust against modifications on the data set by using constraints.

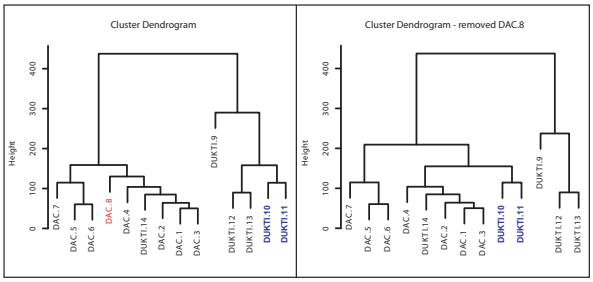

Figure 1: Example for the instability problem in hierarchical cluster analysis. The data is taken from expression profiles of pancreatic ductal adenocarcinoma (DAC) and normal pancreas (DUKTI), see [4]. On the left a dendrogram resulting from a hierarchical cluster analysis of the data set is shown. By removing one sample (DAC.8) the structure of the resulting dendrogram changes markedly.

In recent years various partitional cluster algorithms were adapted to make use of this kind of background information either by constraining the search process or by modifiying the underlying metric. It has been shown that including background knowledge might improve the accuracy of cluster results, i.e. the computed clusters better match a given classification of the data [1, 5, 6, 7, 8, 9, 10]. Basu et al. [11] proposed a probabilistic framework for semi-supervised clustering based on Hidden Markov Random Fields. Kulis et al. [12] demonstrated how to generalize this model to optimize a number of different graph clustering ob-

jectives like ratio cut or normalized cut described by Chan et al. [13] and Shi and Malik [14] respectively. Recently Yan and Domeniconi [15] proposed an extension of this partitional clustering method based on an adaptive kernel method. Davidson et al. [16] proposed a method to include background knowledge in agglomerative hierarchical clustering. In Kestler et al. [3] we analysed the effects of constraints on hierarchical clustering and identified important limitations in the use of background knowledge on hierarchical clustering. We suggested a new semi-supervised divisive hierarchical cluster algorithm which overcomes these restrictions.

In the next Section we present a new semi-supervised divisive hierarchical graph clustering algorithm that uses constraints to influence the clustering process of a weighted undirected graph. The clustering is done by repeatedly removing edges starting from a fully connected graph. Harel and Coren [17] proposed an agglomerative hierarchical clustering method with a preprocessing step based on random walks and showed that a comparison of the paths originating from random walks between adjacent data points induced improved cluster separating edges. We included a similar preprocessing step to reduce the influence of chaining points and to amplify the influence of the background knowledge.

In Section 3 we mention first results from an artificial example as well as from a real world microarray experiment.

2 The Algorithm

The basic clustering task may be summarized as follows. Let $X = \{x_1, \ldots, x_n\}$ be a set of data items with the feature vector $x_l \in \mathbb{R}^d$. X is split into a partition P containing k clusters $p_1, p_2, \ldots, p_k$ by grouping similar objects together such that $\bigcup_{i=1}^{k} p_i = X$, $p_i \neq \emptyset$ and $p_i \cap p_j = \emptyset, i \neq j$. A hierarchical clustering algorithm builds a nested sequence of such partitions.

Algorithm 1 summarises our new semi-supervised divisive hierarchical clustering procedure. In this graph clustering approach we model the data as a weighted undirected graph. Each edge of the fully connected graph is weighted by the distance D of the associated nodes with $D = exp\left(-\frac{d(a,b)^2}{\Delta^2}\right)$, where $d(i,j)$ is the Euclidean distance between two nodes and Δ is the average Euclidean distance between two adjacent nodes in the graph.

Algorithm 1 Semigraster

1. Set up the graph:
   – Build the $N \times N$ adjacency matrix of the fully connected undirected graph representing the data set.
   – Compute the edge weights of the graph, e.g. using a function of the weighted distance between adjacent nodes.
2. Include the background knowledge:
   – Set the edge weights of all edges with cannot-links to 0 and with must-links to Inf.
3. Enhance the edge weights:
   – Compute random walks of lenght k originating from all nodes in X.
   – Compare the paths of random walks from all adjacent nodes by their similarity. Amplify edge weigths between nodes with similar random paths. Reduce edge weights between nodes with differing random paths.
4. Compute the dendrogram:
   – Sequently remove edges indicating a low similarity between adjacent nodes and save emerging clusters.

After initialising the graph structure all background information is added. A must-link indicates that two data items must be in the same cluster and is used by setting the weight of the edge connecting the participating nodes to a predefined maximal value. A cannot-link indicates that two data items should not be in the same cluster and is used by setting the weight to 0.

Clustering a weighted undirected graph by subsequently removing edges with low weights may be hindered by chaining nodes. Like in a single linkage agglomerative clustering a chain of adjacent nodes may connect two distant clusters and thus hinder a splitting of spatially separated clusters. We enhance the elimination of these nodes by applying a preprocessing based on random walks. A random walk on graphs is a stochastic process where a path to all neighbours of a start node is found by randomly choosing a connecting edge. The proba-

bility of a transition between nodes i and j is set to $p_{ij} = \frac{w(i,j)}{d_i}$, where $w(i,j)$ is the weight of the associated edge and d_i is the weighted degree of node i. Then $P^k_{visit}(i)$ is the vector whose j-th entry denotes that a random walk originating at node i will visit node j in its k-th step and $P^{\leq k}_{visit}(i)$ defined by $\sum_{l=1}^{k} P^l_{visit}(i)$ denotes the probabilty of a random walk to visit node j at least in its k-th step. A comparison of two random walks originating from nodes i and j leads to an estimation of the overlap between their neighbourhoods. The neighbourhood of two chaining nodes is less similar than the neighbourhood of two nodes from the same cluster. In order to compare the neighbourhood of two adjacent nodes we evaluate the probability of the random walks of length k starting from i and j to visit almost the same nodes. The neighbourhood similarity of nodes i and j is computed by the cosine correlation between $P^{\leq k}_{visit}(i)$ and $P^{\leq k}_{visit}(j)$. Our algorithm reduces the edge weights between nodes with differing random paths and amplifies those with similar random paths. During this step background knowledge constraining an edge is propagated to the neighbourhood of the constrained nodes and might further enhance the detection of chaining nodes.

After the preprocessing step the clustering is done by repeatedly removing edges between nodes with small neighbourhood similarity. A cluster is separated as soon as all edges connecting two different neighbourhoods are removed. This procedure leads to a natural subsequent decomposition of the graph following the single-linkage cluster paradigm.

3 First Results

For a first look on the applicability of the proposed algorithm we tested its impact on clustering an artificial data set and a real world microarray data set. The artificial data set contains three clusters in a two-dimensional space arranged to form a triangular so a hierarchical clustering is very susceptible to modifications on the data set. The microarray data set is taken from a study on small round blue cell tumours of childhood [18]. It contains neuroblastoma samples (NB), rhabdomyosarcoma samples (RMS), non-Hodgkin lymphoma samples (NHL) and samples from the Ewing family of tumors (EWS). Gene expression data from glass cDNA microarrays containing 6567 genes were used.

To test the robustness of the clustering method we clustered 100 times after randomly removing up to 10% of the data and compared the different cluster results on the top level of the dendrogram by using the Constrained Rand Index [1]. The Rand Index evaluates the number of coincident cluster assignments for each pair of data items in two different cluster results. Because some assignments are fixed due to the given constraints this index is corrected for the background knowledge. Table 1 summarises the results from an analysis of the influence of background knowledge on both data sets. It could be seen that even with a small amount of constraints the stability of the dendrograms is enhanced.

Table 1: Stability of clustering the artificial and childhood cancer data sets after randomly removing and constraining some data items. The table displays a summary of the Constrained Rand Index values from different runs. The percentage of held out and constrained data items is denoted with cut and const.

	median	mean	max
Artificial example			
1% cut, unconst	0.95	0.95	1.00
1% cut, 1% const	1.00	1.00	1.00
10% cut, unconst	0.89	0.76	1.00
10% cut, 1% const	1.00	0.83	1.00
10% cut, 20% const	1.00	0.83	1.00
Childhood cancer			
1% cut, unconst	0.68	0.69	1.00
1% cut, 10% const	0.70	0.74	1.00
10% cut, unconst	0.66	0.69	1.00
10% cut, 10% const	0.70	0.73	1.00

4 Conclusion

Unlike related projects on semi-supervised machine learning we are interested in hierarchical clustering. Our main interest concerns the enhancement of robustness in the hierarchical clustering task. In this paper we described a way to include background knowledge into a hierarchical graph clustering algorithm. Additionally we supported the clustering method by a preprocessing step using random walks. We showed that our approach builds more

stable dendrograms than an unconstrained hierarchical clustering algorithm. All effects of manipulating the graph structure can be traced as every step of the proposed clustering procedure operates on an adjacency matrix. This enables estimating the influence of single constraints during the clustering process.

5 Acknowledgement

This work is supported by the Stifterverband für die Deutsche Wissenschaft (HAK), the German Science Foundation, SFB 518, Project C5 (GP and HAK) and the Graduate School Mathematical Analysis of Evolution, Information and Complexity (JMK).

References

[1] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl. Constrained k-means clustering with background knowledge. In C. E. Brodley and A.P. Danyluk (eds.), *Proceedings of 18th International Conference on Machine Learning*, pp. 577–584, San Francisco, July 2001. Morgan Kaufmann.

[2] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. In *Proceedings of the National Academy of Sciences*, volume 95, pp. 14863–14868. National Academy of Sciences, Washington, 1998.

[3] H.A. Kestler, J.M. Kraus, G. Palm, and F. Schwenker. On the effects of constraints in semi-supervised hierarchical clustering. In F. Schwenker and S. Marinai (eds.), *Artificial Neural Networks in Pattern Recognition*, volume 4087 of *LNAI*, pp. 57–66, Berlin, September 2006. Springer.

[4] M. Buchholz, M. Braun, A. Heidenblut, H.A. Kestler, G. Klöppel, W. Schmiegel, S.A. Hahn, J. Lüttges, and T.M. Gress. Transcriptome analysis of microdissected pancreatic intraepithelial neoplastic lesions. *Oncogene*, 24(44):6626–6636, 2005.

[5] S. Basu, A. Banerjee, and R.J. Mooney. Semi-supervised clustering by seeding. In S. Claude and A.G. Hoffmann (eds.), *Proceedings of 19th International Conference on Machine Learning*, pp. 19–26, San Francisco, July 2002. Morgan Kaufmann.

[6] D. Cohn, R. Caruana, and A. McCallum. Semi-supervised clustering with user feedback. Technical Report 1892, Cornell University, 2003.

[7] E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In S. Becker, S. Thrun, and K. Obermayer (eds.), *Advances in Neural Information Processing Systems 15*, pp. 505–512, Cambridge, October 2003. MIT Press.

[8] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In T. Fawcett and N. Mishra (eds.), *Proceedings of 20th International Conference on Machine Learning*, pp. 11–18, Washington, August 2003. AAAI Press.

[9] M. Bilenko and R.J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In L. Getoor, T.E. Senator, P. Domingos, and C. Faloutsos (eds.), *Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining*, pp. 39–48, New York, August 2003. ACM Press.

[10] M. Bilenko, S. Basu, and R.J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In C.E. Brodley (ed.), *Proceedings of the 21st International Conference on Machine Learning*, pp. 81–88, New York, July 2004. ACM Press.

[11] S. Basu, M. Bilenko, and R. Mooney. A probabilistic framework for semi-supervised clustering. In R. Kohavi, J. Gehrke, W. DuMouchel, and J. Ghosh (eds.), *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 59–68, New York, August 2004. ACM Press.

[12] B. Kulis, S. Basu, I. Dhillon, and R. Mooney. Semi-supervised graph clustering: A kernel approach. In L. De Raed and S. Wrobel (eds.), *Proceedings of the 22nd International Conference on Machine Learning*, pp. 457–464, New York, August 2005. ACM Press.

[13] P.K. Chan, M.D.F. Schlag, and J.Y. Zien. Spectral k-way ratio-cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 13(9):1088 – 1096, September 1994.

[14] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[15] B. Yan and C. Domeniconi. An adaptive kernel method for semi-supervised clustering. In J. Fürnkranz, T. Scheffer, and M. Spiliopoulou (eds.), *17th European Conference on Machine Learning*, volume 4212 of *Lecture Notes in Computer Science*, pp. 521–532, September 2006.

[16] I. Davidson and S.S. Ravi. Agglomerative hierarchical clustering with constraints: Theoretical and empirical results. In A. Jorge, L. Torgo, P. Brazdil, R. Camacho, and J. Gama (eds.), *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, volume 3721 of *LNCS*, pp. 59–70, Berlin, October 2005. Springer.

[17] David Harel and Yehuda Koren. On clustering using random walks. In R. Hariharan, M. Mukund, and V. Vinay (eds.), *21st Conference on Foundations of Software Technology and Theoretical Computer Science*, volume 2245 of *Lecture Notes in Computer Science*, pp. 18–41. Springer, 2001.

[18] J. Khan, J. Wei, M. Ringner, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonesco, C. Peterson, and P. Meltzer. Classification and diagnostic prediction of cancer using gene expression profiling and artificial neural networks. *Nature Medicin*, 6(7):673–679, 2001.