A Detailed List of Constrained Clustering Papers

Ian Davidson and Sugato Basu
U.C. Davis and Google Research

Clustering with constraints is an important recent development in the clustering literature. The addition of constraints allows users to incorporate domain expertise into the clustering process by explicitly specifying what are desirable properties in a clustering solution. This is particularly useful for applications in domains where considerable domain expertise already exists. In this first extensive survey of the field, we discuss the uses, benefits and importantly the problems associated with using constraints. We cover approaches that make use of constraints for partitional and hierarchical algorithms to both enforce the constraints or to learn a distance function from the constraints.

REFERENCES

ABE, N. AND MAMITSUKA, H. 1998. Query learning strategies using boosting and bagging. In Proceedings of
the Fifteenth International Conference on Machine Learning (ICML-98). 1–10.

BANERJEE, A., DHILLON, I., GHOSH, J., AND SRA, S. 2005. Clustering on the unit hypersphere using von
Mises-Fisher distributions. Journal of Machine Learning Research 6, 1345–1382.

BANERJEE, A., MERUGU, S., DHILON, I., AND GHOSH, J. 2005. Clustering with Bregman divergences.
Journal of Machine Learning Research 6, 1705–1749.

BAR-HILLEL, A., HERTZ, T., SHENTAL, N., AND WEINSHALL, D. 2003. Learning distance functions using
equivalence relations. In Proceedings of ICML. Washington, DC, 11–18.

BAR-HILLEL, A., HERTZ, T., SHENTAL, N., AND WEINSHALL, D. 2005. Learning a Mahalanobis metric from
equivalence constraints. Journal of Machine Learning Research 6.

BASU, S., BANERJEE, A., AND MOONEY, R. J. 2002. Semi-supervised clustering by seeding. In Proceedings
of ICML. 19–26.

BASU, S., BANERJEE, A., AND MOONEY, R. J. 2004. Active semi-supervision for pairwise constrained clus-
tering. In Proceedings of SIAM SDM.

BASU, S., BILENKO, M., AND MOONEY, R. J. 2004. A probabilistic framework for semi-supervised clustering.
In Proceedings of ACM SIGKDD. Seattle, WA, 59–68.

BASU, S., DAVIDSON, I., AND WAGSTAFF, K. 2008. Clustering with Constraints: Algorithms, Applications
and Theory. Chapman & Hall/CRC Press Data Mining and Knowledge Discovery Series.

BIE, T. D., MOMMA, M., AND CRISTIANINI, N. 2003. Efficiently learning the metric using side-information.
In Proc. of the 14th International Conference on Algorithmic Learning Theory (ALT2003). Lecture Notes in
Artificial Intelligence, vol. 2842. Springer, 175–189.

BILENKO, M. AND MOONEY, R. J. 2003. Adaptive duplicate detection using learnable string similarity mea-
sures. In Proceedings of ACM SIGKDD. Washington, DC, 39–48.

BLAKE, C. L. AND MERZ, C. J. 1998. UCI Repository of Machine Learning Databases. http://www.ics.uci.edu/∼mlearn/MLRepository.html.
BUNTINE, W. L. 1994. Operations for learning with graphical models. Journal of Artificial Intelligence Research 2, 159–225.
CHANG, H. AND YEUNG, D.-Y. 2004. Locally linear metric adaptation for semi-supervised clustering. In Proceedings of 21st International Conference on Machine Learning (ICML-2004).
COHN, D., CARUANA, R., AND MCCALLUM, A. 2003. Semi-supervised clustering with user feedback. Tech. Rep. TR2003-1892, Cornell University.
DAVIDSON, I., BASU, S., AND WAGSTAFF, K. 2006. In Proceedings of the Tenth European Principles and Practice of KDD (PKDD).
DAVIDSON, I., ESTER, M., AND RAVI, S. S. 2007. Efficient incremental clustering with constraints. In Proceedings of the Thirteen ACM Conference on Data Mining and Knowledge Discovery.
DAVIDSON, I. AND RAVI, S. 2005a. Clustering with constraints: Feasibility issues and the k-means algorithm. In Proceedings of the 2005 SIAM International Conference on Data Mining (SDM-05).
DAVIDSON, I. AND RAVI, S. S. 2005b. Hierarchical clustering with constraints: Theory and practice. In Proceedings of the Nineth European Principles and Practice of KDD (PKDD). 59–70.
DAVIDSON, I. AND RAVI, S. S. 2006. Identifying and generating easy sets of constraints for clustering. In Proceedings of the 21st AAAI Conference.
DAVIDSON, I. AND RAVI, S. S. 2007. Hierarchical clustering with constraints: Theory and practice. Knowledge Discovery and Data Mining 14, 1.
DEMIRIZ, A., BENNETT, K. P., AND EMBRECHTS, M. J. 1999. Semi-supervised clustering using genetic algorithms. In Proceedings of ANNIE. 809–814.
DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. JRSSB 39, 1–38.
DHILLON, I. S., FAN, J., AND GUAN, Y. 2001. Efficient clustering of very large document collections. In Data Mining for Scientific and Engineering Applications. Kluwer Academic Publishers.
DHILLON, I. S. AND GUAN, Y. 2003. Information theoretic clustering of sparse co-occurrence data. In Proceedings of ICDM. 517–521.

$\sum$

EISEN, M. B., SPELLMAN, P. T., BROWN, P. O., AND BOTSTEIN, D. 1998. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, USA 95, 14863–14848.
ELKAN, C. 2003. Using the triangle inequality to accelerate k-means. In ICML.
FREUND, Y. AND SCHAPIRE, R. E. 1996. Experiments with a new boosting algorithm. In Proceedings of the Thirteenth International Conference on Machine Learning (ICML-96), L. Saitta, Ed. Morgan Kaufmann, 148–156.
FREUND, Y., SEUNG, H. S., SHAMIR, E., AND TISHBY, N. 1997. Selective sampling using the query by committee algorithm. Machine Learning 28, 133–168.

GONDEK, D. AND HOFMANN, T. 2004. Non-redundant data clustering. In ICDM ’04: Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM’04). IEEE Computer Society, Washington, DC, USA, 75–82.

HERTZ, T., BAR-HILLEL, A., AND WEINSHALL, D. 2004. Boosting margin based distance functions for clustering. In Proceedings of 21st International Conference on Machine Learning (ICML-2004).

HOCHBAUM, D. S. AND SHMOYS, D. B. 1985. A best possible heuristic for the k-center problem. Mathematics of Operations Research 10(2), 180–184.

HOFMANN, T. AND BUHMANN, J. M. 1998. Active data clustering. In Advances in Neural Information Processing Systems 10.

KEARNS, M., MANSOUR, Y., AND NG, A. Y. 1997. An information-theoretic analysis of hard and soft assignment methods for clustering. In Proceedings of UAI. 282–293.

KLEIN, D., KAMVAR, S. D., AND MANNING, C. D. 2002. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In Proceedings of the Nineteenth International Conference on Machine Learning.

KLEINBERG, J. AND TARDOS, E. 1999. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and Markov random fields. In Proceedings of FOCS. 14–23.

LANGE, T., LAW, M. H. C., JAIN, A. K., AND BUHMANN, J. M. 2005. Learning with constrained and unlabeled data. In CVPR. San Diego, CA, 731–738.

LAW, M. H. C., TOPCHY, A., AND JAIN, A. K. 2005. Model-based clustering with probabilistic constraints. In Proceedings of the 2005 SIAM International Conference on Data Mining (SDM-05).

LEWIS, D. AND GALE, W. 1994. A sequential algorithm for training text classifiers. In Proceedings of Seventeenth International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-94).

LU, Z. AND LEEN, T. K. 2005. Semi-supervised learning with penalized probabilistic clustering. In Advances in Neural Information Processing Systems 17.

MCCALLUM, A. AND NIGAM, K. 1998. Employing EM and pool-based active learning for text classification. In Proceedings of ICML. Madison, WI.

NANNI, M. 2005. Speeding-up hierarchical agglomerative clustering in presence of expensive metrics. In PAKDD.

NEAL, R. M. AND HINTON, G. E. 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants. In Learning in Graphical Models, M. I. Jordan, Ed. MIT Press, 355–368.

PELLEG, D. AND BARAS, D. 2007. K-means with large and noisy constraint sets. In ECML.

RAND, W. M. 1971. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66, 366.

SEEGER, M. 2000. Learning with labeled and unlabeled data.

SEGAL, E., WANG, H., AND KOLLER, D. 2003. Discovering molecular pathways from protein interaction and gene expression data. Bioinformatics 19, i264–i272.

SINKKONEN, J. AND KASKI, S. 2000. Semisupervised clustering based on conditional distributions in an auxiliary space. Tech. Rep. A60, Helsinki University of Technology.

WAGSTAFF, K. AND CARDIE, C. 2000. Clustering with instance-level constraints. In Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann, Palo Alto, CA, 1103–1110.

WAGSTAFF, K., CARDIE, C., ROGERS, S., AND SCHROEDL, S. 2001a. Constrained k-means clustering with background knowledge. In Proceedings of the Eighteenth International Conference on Machine Learning.

WAGSTAFF, K., CARDIE, C., ROGERS, S., AND SCHROEDL, S. 2001b. Constrained K-Means clustering with background knowledge. In Proceedings of ICML. 577–584.

$\sum$

WAGSTAFF, K. L. 2002. Intelligent Clustering with Instance-Level Constraints. Ph.D. thesis, Cornell University.

XENARIOS, I., FERNANDEZ, E., SALWINSKI, L., DUAN, X. J., THOMPSON, M. J., MARCOTTE, E. M., AND EISENBERG, D. 2001. DIP: The database of interacting proteins: 2001 update. Nucleic Acids Research 29, 1, 239–241.

XING, E. P., NG, A. Y., JORDAN, M. I., AND RUSSELL, S. 2003. Distance metric learning, with application to clustering with side-information. In NIPS 15.

YAN, R., ZHANG, J., YANG, J., AND HAUPTMANN, A. G. 2004. A discriminative learning framework with pairwise constraints for video object classification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).

YANG, K., YANG, M., AND KAFATOS, M. 2001. A feasible method to find areas with constraints using hierarchical depth-first clustering. In Scientific and Statistical Database Management Conference.

ZAIANE, O., FOSS, A., LEE, C., AND WANG, W. 2000. On data clustering analysis: Scalability, constraints and validation. In PAKDD.

ZHU, X. 2005. Semi-supervised learning literature survey. Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison. http://www.cs.wisc.edu/∼jerryzhu/pub/ssl survey.pdf.