Proceedings of the Ninth AAAI/SIGART Doctoral Consortium,
pp. 979-980, San Jose, CA, July 2004

Semi-supervised Clustering with Limited Background Knowledge

Sugato Basu
Email: sugato@cs.utexas.edu
Address: Department of Computer Sciences, University of Texas at Austin, Austin, TX - 78712, USA

Thesis Goal

In many machine learning domains, there is a large supply of unlabeled data but limited labeled data, which can be expensive to generate. Consequently, semi-supervised learning, learning from a combination of both labeled and unlabeled data, has become a topic of significant recent interest. Our research focus is on semi-supervised clustering, which uses a small amount of supervised data in the form of class labels or pairwise constraints on some examples to aid unsupervised clustering. Semi-supervised clustering can be either constraint-based, i.e., changes are made to the clustering objective to satisfy user-specified labels/constraints, or metric-based, i.e., the clustering distortion measure is trained to satisfy the given labels/constraints. Our main goal in this thesis is to study constraint-based semi-supervised clustering algorithms, integrate them with metric-based approaches, characterize some of their properties and empirically validate our algorithms on different domains, e.g., text processing and bioinformatics.

Background

Existing methods for semi-supervised clustering fall into two general approaches that we call *constraint-based* and *metric-based* methods.

In constraint-based approaches, the clustering algorithm itself is modified so that user-provided labels or constraints are used to get a more appropriate clustering. Previous work in this area includes modifying the clustering objective function so that it includes a term for satisfying specified constraints (Demiriz, Bennett, & Embrechts 1999), and enforcing constraints to be satisfied during the cluster assignment in the clustering process (Wagstaff *et al.* 2001).

In metric-based approaches, an existing clustering algorithm that uses a particular distortion measure is employed; however, the measure is first trained to satisfy the labels or constraints in the supervised data. Several distortion measures have been used for metric-based semi-supervised clustering, including Jensen-Shannon divergence trained using gradient descent (Cohn, Caruana, & McCallum 2003), Euclidean distance modified by a shortest-path algorithm (Klein, Kamvar, & Manning 2002), or Mahalanobis distances trained using convex optimization (Bar-Hillel *et al.* 2003; Xing *et al.* 2003).

However, metric-based and constraint-based approaches to semi-supervised clustering have not been adequately compared in previous work, and so their relative strengths and weaknesses are largely unknown.

An important domain that motivates the semi-supervised clustering problem is the clustering of genes for functional prediction. For most organisms, only a limited number of genes are annotated with their functional pathways, with the majority of the genes still having unknown functions. Categorization of these genes into functional groups using gene microarray data, phylogenetic profiles, etc. is a natural semi-supervised clustering problem. Clustering (with model selection to choose the right number of clusters) is more well suited to this domain than classification, since the number of functional classes is not known a priori. Moreover background knowledge, available in the form of functional pathway labels (KEGG, GO) or constraints over some of the genes (DIP), could easily be incorporated as supervision to improve the clustering accuracy.

Progress

In our first work, we showed how supervision in the form of labeled data can be incorporated into clustering (Basu, Banerjee, & Mooney 2002). The labeled data were used to generate seed clusters for initializing model-based clustering algorithms, and constraints generated from the labeled data were used to guide the clustering process towards a partitioning similar to the user-specified labels. We showed that the K-Means algorithm is equivalent to an EM algorithm on a mixture of K Gaussians under assumptions of identity covariance of the Gaussians, uniform mixture component priors and expectation under a particular type of conditional distribution. This underlying model helps us to prove convergence guarantees for the proposed label-based semi-supervised clustering algorithms.

Next, we showed that semi-supervised clustering with pairwise *must-link* and *cannot-link* constraints has an underlying probabilistic model – a Hidden Markov Random Field (HMRF) (Basu, Banerjee, & Mooney 2004). In this work, we also outlined a method for selecting maximally informative constraints in a query-driven framework for pairwise constrained clustering. In order to maximize the utility of

the limited supervised data available in a semi-supervised setting, supervised training examples should be, if possible, *actively* selected as maximally informative ones rather than chosen at random. This would imply that fewer constraints will be required to significantly improve the clustering accuracy. To this end, a new algorithm was developed to actively select good pairwise constraints for semi-supervised clustering, using an active learning strategy based on farthest-first traversal. The proposed scheme has two phases: (a) explore the given data to get pairwise disjoint non-null neighborhoods, each belonging to a different cluster in the underlying true categorization of the data, within a small number of queries, and (b) consolidate this cluster structure using the remaining queries, to get better centroid estimates.

In recent work, we have shown that the HMRF clustering model is able to incorporate any Bregman divergence (Banerjee *et al.* 2004) as the clustering distortion measure, which allows using the framework with such common distortion measures as KL-divergence, I-divergence, and parameterized squared Mahalanobis distance. Additionally, cosine similarity can also be used as the clustering distortion measure in the framework, which makes it useful for directional datasets (Basu, Bilenko, & Mooney 2004). For all such measures, minimizing the semi-supervised clustering objective function becomes equivalent to finding the maximum *a posteriori* probability (MAP) configuration of the underlying HMRF.

We have also developed a new semi-supervised clustering approach that unifies constraint-based and metric-based techniques in an integrated framework (Bilenko, Basu, & Mooney 2004). This algorithm trains the distortion measure with each clustering iteration, utilizing both unlabeled data and pairwise constraints. The formulation is able to learn individual metrics for each cluster, which permits clusters of different shapes. This work also explores metric learning for feature generation (in contrast to simple feature weighting), which we empirically demonstrate to outperform current state-of-the-art metric learning algorithms, under certain conditions.

In all these projects, experiments have been performed on both low dimensional UCI datasets and high dimensional text data sets, using KMeans and EM as the baseline clustering algorithms.

Proposed Research

In future, we want to study the following aspects of semi-supervised clustering, in decreasing order of priority:

(1) Semi-supervised approaches for finding *overlapping* clusters in the data. This is especially relevant for gene clustering in the bioinformatics domain, since genes often belong to multiple functional pathways.

(2) The feasibility of semi-supervising other clustering algorithms, e.g., spectral clustering, agglomerative clustering, etc. We are currently exploring semi-supervised versions of kernel-based clustering, which would be useful for datasets that are not linearly separable.

(3) Application of the semi-supervised clustering model to other domains apart from UCI datasets and text. We are currently focusing on two domains: (a) search result clustering of web search engines, e.g., Google, and (b) clustering of gene microarray data in bioinformatics.

(4) Effect of noisy or probabilistic supervision in pairwise constrained clustering. This study will be especially important for deploying our proposed semi-supervised clustering algorithms to practical settings, where background knowledge would be in general noisy.

(5) Model selection using both unsupervised data and the limited supervised data, for automatic selection of number of clusters in semi-supervised clustering. Most model selection criteria in clustering are only based on unsupervised data – we want to explore whether supervised data available in the form of labels or constraints can be used to select the number of clusters more effectively.

(6) Theoretical study of the relative benefits of supervised and unsupervised data in semi-supervised clustering, similar to the analysis of (Ratsaby & Venkatesh 1995).

References

Banerjee, A.; Merugu, S.; Dhillon, I. S.; and Ghosh, J. 2004. Clustering with Bregman divergences. In *Proc. of the 2004 SIAM Intl. Conf. on Data Mining (SDM-04)*.

Bar-Hillel, A.; Hertz, T.; Shental, N.; and Weinshall, D. 2003. Learning distance functions using equivalence relations. In *Proc. of 20th Intl. Conf. on Machine Learning (ICML-2003)*, 11–18.

Basu, S.; Banerjee, A.; and Mooney, R. J. 2002. Semi-supervised clustering by seeding. In *Proc. of 19th Intl. Conf. on Machine Learning (ICML-2002)*, 19–26.

Basu, S.; Banerjee, A.; and Mooney, R. J. 2004. Active semi-supervision for pairwise constrained clustering. In *Proc. of the 2004 SIAM Intl. Conf. on Data Mining (SDM-04)*.

Basu, S.; Bilenko, M.; and Mooney, R. J. 2004. A probabilistic framework for semi-supervised clustering. In submission, available at `http://www.cs.utexas.edu/~ml/publication`.

Bilenko, M.; Basu, S.; and Mooney, R. J. 2004. Integrating constraints and metric learning in semi-supervised clustering. In *Proc. of 21st Intl. Conf. on Machine Learning (ICML-2004)*.

Cohn, D.; Caruana, R.; and McCallum, A. 2003. Semi-supervised clustering with user feedback. Technical Report TR2003-1892, Cornell University.

Demiriz, A.; Bennett, K. P.; and Embrechts, M. J. 1999. Semi-supervised clustering using genetic algorithms. In *Artificial Neural Networks in Engineering (ANNIE-99)*, 809–814.

Klein, D.; Kamvar, S. D.; and Manning, C. 2002. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *Proc. of 19th Intl. Conf. on Machine Learning (ICML-2002)*, 307–314.

Ratsaby, J., and Venkatesh, S. S. 1995. Learning from a mixture of labeled and unlabeled examples with parametric side information. In *Proc. of the 8th Annual Conf. on Computational Learning Theory*, 412–417.

Wagstaff, K.; Cardie, C.; Rogers, S.; and Schroedl, S. 2001. Constrained K-Means clustering with background knowledge. In *Proc. of 18th Intl. Conf. on Machine Learning (ICML-2001)*, 577–584.

Xing, E. P.; Ng, A. Y.; Jordan, M. I.; and Russell, S. 2003. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, 505–512. Cambridge, MA: MIT Press.