Proceedings of the Ninth AAAI/SIGART Doctoral Consortium,
pp. 981-982, San Jose, CA, July 2004

Learnable Similarity Functions and Their Applications to Clustering and Record Linkage

Mikhail Bilenko
Department of Computer Sciences
University of Texas at Austin
1 University Station C0500
Austin, TX 78712-1188, USA
mbilenko@cs.utexas.edu

Many problems in machine learning and data mining depend on distance estimates between observations, e.g., instance-based classification, clustering, information retrieval, and record linkage in databases. However, the appropriate notion of similarity can vary depending on the particular domain, dataset, or task at hand. Consequently, a large number of functions that compute similarity between objects have been developed for different data types, varying greatly in their expressiveness, mathematical properties, and assumptions (Gusfield 1997; Duda, Hart, & Stork 2001). Additionally, there exists a substantial body of research on feature space transformations that attempt to provide a more salient representation of data than the original feature space, e.g. Principal Component Analysis (Jolliffe 1986) and Locally Linear Embedding (Saul & Roweis 2003).

All of these techniques make certain assumptions about the optimal representation of data and its influence on computing similarity which may or may not be applicable for specific datasets and tasks. Therefore, it is desirable to learn accurate similarity functions from training data to capture the correct notion of distance for a particular task at hand in a given domain. Recently, several approaches have been suggested for training such functions using *pairwise relations* between instances, e.g. pairwise equivalence (Ristad & Yianilos 1998; Cohen & Richman 2002), common cluster membership (Xing *et al.* 2003), and relative comparisons (Schultz & Joachims 2004). These approaches have shown improvements over traditional similarity functions for different data types such as vectors in Euclidean space, strings, and database records composed of multiple text fields. While these initial results are encouraging, there still remains a large number of similarity functions that are currently unable to adapt to a particular domain. In our research, we attempt to bridge this gap by developing both *new learnable similarity functions* and methods for their *application to particular problems* in machine learning and data mining.

In preliminary work, we proposed two learnable similarity functions for strings that adapt distance computations given training pairs of equivalent and non-equivalent strings (Bilenko & Mooney 2003a). The first function is based on a probabilistic model of edit distance with affine gaps (Gusfield 1997), a widely used character-based metric. The second function is a variant of cosine similarity that utilizes a Support Vector Machine (SVM) classifier (Vapnik 1998) to discriminate between similar and dissimilar string pairs. As opposed to their *static* analogs, these similarity functions *adapt* to a particular domain using training examples and produce more accurate similarity estimates as a result.

These string similarity functions were inspired by the *record linkage* problem, which is the task of identifying semantically equivalent database records that are syntactically different. Record linkage is one of the key tasks in data cleaning since presence of unidentified duplicate records violates data integrity principles. Information integration from multiple sources also relies on record linkage methods because information describing the same entity must be linked across sources. We developed a system, MARLIN (Multiply Adaptive Record Linkage using INduction), that we used as a testbed for evaluating learnable string similarity measures on the record linkage task. MARLIN incorporates learnable string metrics in a two-layer learning framework where multiple string distances are combined to identify duplicate database records. MARLIN was used to experimentally demonstrate that learnable string distances improve over traditional distance functions on a number of real-world record linkage datasets (Bilenko & Mooney 2003a).

Prior research on active learning has demonstrated that the learning process can be facilitated by intelligent selection of informative training examples from a pool of unlabeled data. Because training data for learnable similarity functions is typically composed of object pairs, selecting meaningful training examples presents unique challenges that are distinct from traditional active learning methods in classification. Based on initial experiments with MARLIN, we proposed two strategies for selectively collecting pairwise training data, static-active sampling and weakly-labeled selection, that facilitate training of adaptive similarity functions for the record linkage task (Bilenko & Mooney 2003b).

Another important application that can benefit from using learnable similarity functions is clustering. While traditionally clustering has been viewed as an unsupervised learning problem, recently there has been increasing attention to *semi-supervised clustering*, where limited supervision is provided to obtain better grouping of data (Wagstaff *et al.* 2001; Xing *et al.* 2003). In initial work on employing learnable distance functions in clustering, we developed the MPCK-MEANS algorithm, which is a semi-supervised variant of unsupervised K-means clustering. MPCK-MEANS utilizes training data in the form of pairwise constraints in a unified framework that encompasses cluster initialization, constraint satisfaction, and learning individual parameterized Mahalanobis distances for each cluster. Our initial results have shown that using both distance metric learning and cluster seeding with constraints leads to improvements over unsupervised clustering as well as over the individual methods in isolation (Bilenko, Basu, & Mooney 2004). In recent work, we have presented a probabilistic framework for semi-supervised clustering, HMRF-KMEANS, which generalizes MPCK-MEANS to other distance measures including Bregman divergences and cosine similarity while presenting a view of semi-supervised clustering as the task of label assignment in Hidden Markov Random Fields (Basu, Bilenko, & Mooney 2004).

While these initial results are encouraging, there is a number of open avenues for future research in developing and applying learnable similarity functions. Following are several directions that appear most promising to pursue in the near future.

We plan to extend the vector-space string similarity based on pairwise classification with SVMs by incorporating alternative vector-space representations that go beyond simple tokenization, e.g. string subsequence kernels (Lodhi *et al.* 2002). Additionally, we intend to modify SVM training methods to avoid overly aggressive feature selection performed by traditional SVM algorithms.

While our initial learnable model for string edit distance with affine gaps uses Hidden Markov Models in a generative framework, we will attempt to model string edit distance using undirected graphical models such as Conditional Random Fields (Lafferty, McCallum, & Pereira 2001). Because of the greater representational power of such models, we hope that this research will lead to edit distances that go beyond sequential character alignment and incorporate long-range dependencies.

Finally, we hope to extend usage of learnable similarity measures beyond record linkage to related information integration tasks, information extraction and schema mapping (Doan, Domingos, & Halevy 2001), and combine these tasks in a single cohesive framework. Because all of these tasks involve identifying equivalent objects represented differently, employing adaptive similarity functions in a unified framework could improve accuracy on the individual tasks as well as for the overall information integration process.

We hope that progress in these directions will lead to new learnable similarity functions that will advance the state-of-the-art in clustering and information integration. We believe that usefulness of learnable distance measures stretches far beyond the applications outlined here, and this research will lay the groundwork for developing and applying learnable similarity functions in other areas such as vision and information retrieval.

References

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., and Mooney, R. J. 2003a. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003)*, 39–48.

Bilenko, M., and Mooney, R. J. 2003b. On evaluation and training-set construction for duplicate detection. In *Proceedings of the KDD-03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, 7–12.

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

Cohen, W. W., and Richman, J. 2002. Learning to match and cluster large high-dimensional data sets for data integration. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*.

Doan, A.; Domingos, P.; and Halevy, A. 2001. Reconciling schemas of disparate data sources: A machine-learning approach. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, 509–520. Santa Barbara, CA: ACM Press.

Duda, R. O.; Hart, P. E.; and Stork, D. G. 2001. *Pattern Classification*. New York: Wiley, second edition.

Gusfield, D. 1997. *Algorithms on Strings, Trees and Sequences*. New York: Cambridge University Press.

Jolliffe, I. T. 1986. *Principal Component Analysis*. New York: Springer-Verlag.

Lafferty, J.; McCallum, A.; and Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of 18th International Conference on Machine Learning (ICML-2001)*.

Lodhi, H.; Saunders, C.; Shawe-Taylor, J.; Cristianini, N.; and Watkins, C. 2002. Text classification using string kernels. *Journal of Machine Learning Research* 2:419–444.

Ristad, E. S., and Yianilos, P. N. 1998. Learning string edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(5):522–532.

Saul, L., and Roweis, S. 2003. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research* 4:119–155.

Schultz, M., and Joachims, T. 2004. Learning a distance metric from relative comparisons. In *Advances in Neural Information Processing Systems 16*.

Vapnik, V. N. 1998. *Statistical Learning Theory*. John Wiley & Sons.

Wagstaff, K.; Cardie, C.; Rogers, S.; and Schroedl, S. 2001. Constrained K-Means clustering with background knowledge. In *Proceedings of 18th International Conference 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.