A Bayesian Approach to Similarity with Applications to OCR and 3D Object Recognition

Thomas M. Breuel
PARC, Palo Alto, USA

January 18, 2003

Judging similarity is a key human skill and at the heart of many pattern recognition methods: good models of similarity allow us to extrapolate to novel data from fewer samples. Similarity has been studied extensively in areas such as pattern recognition and adaptive nearest neighbors [10, 6, 1], character recognition, text-based information retrieval [4, 7], and case-based reasoning [5]. Closely related to models of similarity is the problem of multitask learning [3], where we attempt to use information learned from one task to improve performance on a related task.

In this talk, I explore the relationship between Bayes optimal classification, similarity measures, and multitask learning. The key for making the connection is the conditional distribution $P(S|x,x')$, where S is a boolean variable indicating whether samples x and x' come from the same class.

One of the interests in using $P(S|x,x')$ for determining similarity is that it can be viewed as a two-class classification problem using the single concatenated feature vector (x,x'). This means that a large number of known classification methods (logistic regression, multilayer perceptrons, hidden Markov models, etc.) can be used for modeling $P(S|x,x')$; previous approaches to pairwise probabilities (like [5, 8]) have relied on joint probability distributions, attempting to model $P(x,x'|S)$. Another practical advantage is that training data for $P(S|x,x')$ does not require class membership to be known–it only requires us to be able to identify samples known to come from the same class. This allows us to train $P(S|x,x')$ in some situations (e.g., 3D object recognition) in which hand-labeled data is unavailable.

I review and prove connections of $P(S|x,x')$ with posterior densities $P(\omega|x)$. First, I show that under simple assumptions, classification with $P(S|x,x')$ and a set of prototypes is Bayes-optimal and hence equivalent to classification using posterior densities $P(\omega|x)$. Furthermore, even if $P(S|x,x')$ is not estimated correctly, as long as the estimate is sufficiently smooth, it has the same asymptotic performance guarantees as Euclidean nearest neighbor methods. These properties suggest that we do not lose classification performance by using $P(S|x,x')$ instead of nearest neighbor classification methods or classifiers based on estimates of the posterior densities.

A particularly important class of classification problems are those in which samples for each class are generated from prototypes together with a noise model (dependent or independent of the prototypes). Many OCR and object recognition problems are at least approximately of this form. I show that learning $P(S|x,x')$ allows considerable transfer of knowledge between different problem instances. Furthermore, for the case of Gaussian noise, I show direct equivalence to nearest neighbor methods using a quadratic form.

I describe applications to handwriting recognition and OCR problems [2, 11]. In these applications, font, image degradation, or writer identity give rise to classes of related classification problems. I show how collections of prototypes can be recovered using nearest neighbor methods and hierarchical Bayesian methods. I demonstrate severalfold improvements in error rates relative to standard classification methods (mixture discriminant analysis, multilayer perceptrons, style modeling, hierarchical mixtures of experts).

Another important application area is the problem of 3D generalization in object recognition. Most previous methods for learning 3D object models, or for generalizing from given views, have required multiple training or sample views per object (e.g., [13, 12]). However, human observers show significant ability to generalize from a single view of a previously unseen object to novel views (single view generalization). It has been argued in the literature [9] that combinations of multiple views are required to account for human recognition performance. I demonstrate that learning of Bayesian similarity models results in single view generalization that exceeds that previously demonstrated in the literature (error rates are reduced severalfold), both on simulated wireframe objects and on a published database of real images (COIL-100). I also show how models of $P(S|x,x')$ for 3D object recognition can be acquired in an unsupervised manner from motion sequences and demonstrate object class-specific recognition effects similar to those observed in humans.

Overall, these results show that the use of $P(S|x,x')$ give us a powerful tool for modeling similarity: classifiers based on Bayesian statistical similarity provide solutions to multitask classification problems, can be acquired from unlabeled training data in some cases, and provide novel models of 3D single view generalization.

References

[1] Jonathan Baxter. The canonical distortion measure for vector quantization and function approximation. In *Proc. 14th International Conference on Machine Learning*, pages 39–47. Morgan Kaufmann, 1997.

[2] T.M. Breuel. Classification by probabilistic clustering. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (ICASSP 2001)*, pages 1333–1336, 2001.

[3] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[4] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 1990.

[5] Boi Faltings. Probabilistic indexing for case-based prediction. In *ICCBR*, pages 611–622, 1997.

[6] Trevor Hastie and Robert Tibshirani. Discriminant adaptive nearest neighbor classification and regression. In David S. Touretzky, Michael C. Mozer, and Michael E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 409–415. The MIT Press, 1996.

[7] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pages 50–57, Berkeley, California, August 1999.

[8] Thomas Hofmann, Jan Puzicha, and Michael I. Jordan. Learning from dyadic data. In *Advances in Neural Information Processing Systems (NIPS'98)*, pages 466–472, 1999.

[9] Z. Liu and D. Kersten. 2D observers for human 3D object recognition? *Vision Research*, 38:2507–2519, 1998.

[10] David G. Lowe. Similarity metric learning for a variable-kernel classifier. *Neural Computation*, 7(1):72–85, 1995.

[11] C. Mathis and T. M. Breuel. Classification using a hierarchical bayesian approach. In *Proceedings of the International Conference on Pattern Recognition (ICPR'02), Quebec City, Quebec, Canada*, 2002.

[12] T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266, 1990.

[13] Shimon Ullman and Ronen Basri. Recognition by Linear Combinations of Models. A.I. Memo No. 1152, MIT Artificial Intelligence Laboratory, Cambridge, MA, USA, 1989.