Applying Clustering to the Classification Problem

Piew Datta
Department of Information and Computer Science
University of California
Irvine, CA 92717
pdatta@ics.uci.edu

The minimum-distance classifier (Duda & Hart, 1973) learns a single mean prototype for each class and uses a nearest neighbor approach for classification. A problem arises when classes cannot be accurately represented using a single prototype; multiple prototypes may be necessary. Our approach is to find groups of examples for each of the classes, generalize these groups into prototypes using a mean representation, and then classify using a nearest neighbor approach.

K-means clustering is applied in unsupervised environments for finding groupings of examples. The problem with k-means clustering is finding the correct number of groups, k_c for each class, c (Duda & Hart, 1973; Smyth,1996; Cheeseman & Stutz,1996). Although our task is in a supervised environment, k-means clustering can still be applied. We propose three methods for finding k_c in the supervised classification task.

The first method, named SNMJ, consists of calculating the sum squared error of k_c clusters and k_c+1 clusters (Duda & Hart, 1973, p. 242). The sum of squared error for k_c is defined by $J_e(k_c) = \sum_{i=1}^{k_c} \sum_{x \in X_i} |x - m_i|^2$, where X_i is cluster i and m_i is the mean for a particular attribute in cluster i. Since J_e will always decrease as k_c increases, we compare the ratio of $J_e(k_c+1)$ to $J_e(k_c)$. If there is a statistically significant improvement in $J_e(k_c+1)$ then we would consider having k_c+1 clusters instead of k_c. The threshold Duda & Hart use is $threshold = 1 - \frac{2}{\pi d} - \alpha\sqrt{\frac{2(1-8/\pi^2 d)}{nd}}$, where d is the number of attributes, n is the number of examples, and α is the degree of significance.

The second method, named SNMC, uses the training set accuracy and hill-climbing search to determine k_c. Initially each k_c is set to 1. SNMC iteratively attempts to increment by one the number of clusters in each class sequentially, checking the classification accuracy on the training set after each attempt, and increasing k_c if an increase in training set accuracy occurs.

The last method, named SNMC-D, attempts to increase the distance among prototypes of differing classes by incrementing the number of prototypes in a class. Intuitively, we want the prototypes for the classes to be as dissimilar as possible. Figure 1a. shows that prototypes P and N have a small distance between them. By increasing the number of prototypes for +, the distance between the prototypes has increased, resulting in a better representation of the examples (Figure. 1b.).

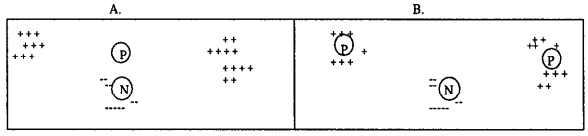

Figure1: Increasing the number of prototypes for + increases the distance between prototypes of different classes.

We experimentally evaluated SNMJ and SNMC comparing their average classification accuracy on 20 domains from the UCI data repository (Murphy & Aha, 1994). We ran them 30 times on each domain with 66.6% training data and 33.3% test data. The average accuracy for SNMJ is 78.9, and for SNMC is 82.8. This shows that SNMC classifies better than SNMJ on this set of domains. Refer to Datta & Kibler(1997) for more detailed results on SNMC.

Acknowledgments

We would like to thank Dennis Kibler and Pedro Domingos for providing fruitful discussions on this research.

References

Cheeseman, P. & Stutz, P. (1988). AutoClass: A Bayesian Classification System. In Proceedings of the Fifth International Conference on Machine Learning. Morgan Kaufmann.

Datta, P. and Kibler, D. (1997). Symbolic Nearest Mean Classifiers. In Proceedings of the National Conference on Artificial Intelligence. Providence, RI. AAAI/MIT Press.

Duda, R., and Hart P. (1973). *Pattern classification and scene analysis.* New York: John Wiley & Sons.

Murphy, P. and Aha, D. (1994). UCI repository of machine learning databases [machine readable data repository]. Tech. Rep., University of California, Irvine.

Smyth, P. (1996). Clustering using Monte Carlo Cross-Validation. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining.* Seattle, Washington.