Improving Classification Accuracy By Learning Multiple Category Prototypes

Mukesh Rohatgi (MXR100F@ECONOMY.BPA.ODU.EDU)
IS/DS Department, Old Dominion University
Norfolk, VA 23529, USA

Introduction

The purpose of this paper is to demonstrate that the accuracy of classifier systems using prototype-based category descriptions can be improved by learning multiple category prototypes. Learning of multiple prototypes is especially useful when we encounter highly variable categories or come across overlapping contrast categories. Under the circumstances just described it becomes infeasible to accurately classify object instances on the basis of a single prototype representing the central tendency of category instances. This happens because the instance to be classified is either too dissimilar from the category prototype or is much more similar to the prototypes of contrasting categories. A feasible solution in such cases is to learn exemplars that typically get mapped towards the category boundary and use them as the basis for accurate classification. If a substantial number of such exemplars exhibit similar characteristics then it becomes possible to generate an abstraction-based representation for these outliers. Such representations can be assumed to capture the central tendency of a subclass contained within a highly variable category.

The value of learning class and subclass information in the form of multiple prototypes is quite evident. During classification, a test instance which is substantially dissimilar to the central category prototype may turn out to be similar to a prototype representing a subclass of the category. For such cases we can claim that the test instance belongs to the parent of the subclass. Therefore, with multiple prototypes, conclusions about category membership of a classification instance have a greater probability of being correct when compared to conclusions based on a single category prototype.

Multiple Prototype Learning Process

The following summarizes the multiple prototype learning process used by a computer-based system implemented as a part of this research (for details refer to Rohatgi, 1994):

1. Acquire object descriptions in terms of attribute-value pairs.
2. Generate a single prototype-based representation for each object class present in the training data set. The prototype-based representation of a category is generated by abstracting the central tendency of category instances described to the system during the first step.
3. Generate additional prototypes for each object class with the help of a failure-based classification process. Additional category prototypes generated by the system can be understood as abstractions representing the subclasses of a highly variable class. A subclass prototype is based on training set instances that get inaccurately classified by existing category prototypes. Each failure serves as a seed for the generation of an additional prototype representing a new subclass.
4. Identify the optimal multiple prototype representation through iterative elimination of prototypes with a low degree of abstraction. The elimination process is controlled with the help of a truncation parameter.

Evaluation

Classification experiments (for details see Rohatgi, 1994) with the iris and breast cancer data downloaded from the machine learning repository available on the Internet node ics.uci.edu show that under certain circumstances classification accuracy can be improved by choosing an optimum number of multiple prototypes to represent a category. For example, classification accuracy in the case of breast cancer data improved from 0.60 to 0.73 by moving from a single prototype to multiple prototypes.

Circumstances that favor the multiple prototype approach seem to involve cases with overlapping categories. In a case involving well separated categories (similar to the iris data), the approach suggested in this paper may not prove to be fruitful or may not be even required. Whenever applicable, the approach requires the selection of a suitable value for the truncation parameter (0.20 in the case of breast cancer data). The function of the truncation parameter is to optimize the number of prototypes used for representing target classes during classification.

Future Work

The conclusions outlined in the previous section are based on a limited set of experiments. Further experimentation with the prototype system is required to substantiate the validity of our results. In the future we plan to conduct few more experiments with the prototype to explore its classification behavior in other types of artificial and natural domains. In particular, the plan is to direct our future effort towards the type of experimentation that will help us in characterizing the role of truncation parameter in classification tasks.

References

Rohatgi, M. (1994). *A human Learning Approach For Designing Adaptive Knowledge-Based Systems*. Doctoral Dissertation, IS Dept., Texas Tech Univ., Lubbock, Texas.