A Comparison of Methods for the Reduction of Attributes before Classification in Data Mining

Eleanor E. WIESCHWAUS and Martin SCHULTZ, advisor

Abstract-- The problem of classification has been well examined when the number of variables is small, say, fifty or fewer. Less work has been done as the number of variables increases. Many algorithms shown to be accurate and efficient with few features become noisy and imprecise as the number grows towards the thousands. This has ramifications in the application of many classification problems. Imagine that a skin sample taken from a patient is to determine if he or she has cancer. If the diagnostician does not know which the relevant elements are, he or she can test for thousands. The results of these tests can then be cross-referenced with a database of other patient’s variable levels and their ultimate diagnoses. The question that arises is the most efficient way to organize the data such that a diagnosis can be made quickly. Machine learning techniques like decision trees, neural networks, and nearest neighbor can be employed to organize data such that classification is efficient and accurate when the number of variables is small. The purpose of this project is to study methods for classification when the number of features is quite large. Generally, we found that the most effective classification methods involved several levels. In the first and second level, the number of attributes was reduced, and in the final stage, one of the previously mentioned standards was applied to the reduced set. Of the variable reduction algorithms tested, those that involved elements of randomization and evolution fared best in selecting relevant attributes, particularly when populations were seeded in the second stage with the result from the first.