Discernibility-based Algorithms for Classification

Zacharias Voulgaris[a], and George D. Magoulas [b],
[a]Department of Electrical and Computer Engineering,
Georgia Institute of Technology, Atlanta, GA, U.S.A.
[b]Department of Computer Science and Information Systems,
Birkbeck College, University of London, U.K.
zvoulgaris@gatech.edu,gmagoulas@dcs.bbk.ac.uk

Key words: Classification, pattern recognition, discernibility, ensemble, feature resampling, k-nearest neighbor, distance metrics

There has been a great deal of research on classification systems, aiming at the development of general methods and specialised techniques to tackle particular classification problems. These techniques often employ the statistical properties of the data involved, or adaptive mechanisms to exploit every little piece of information that may reveal a useful property of the data, leading to a reliable classification. Yet, datasets not always follow the assumed statistical distribution and many classifiers tend to become "confused" when dealing with large datasets, as the excessive information that is there often compromises the classifiers' performance, sometimes due to the problem of overfitting. This can be attributed to the fact that they often consider all patterns being equally important, instead of taking into account their underlying structure which may yield more useful information for the classification process.

In this paper we propose an approach to enhance the pattern classification process by taking into account the geometry of class structure in datasets of interest. This is based on the recently proposed Discernibility concept, through one of its indexes, namely the Spherical Index of Discernibility. We demonstrate how this index can be applied to enhance distance-based classifiers, such as the k-nearest neighbor, as well as information fusion in diverse classifier ensembles. The ensemble used in our work comprised of a number of different classifiers in order to ensure a high level of diversity in the errors of the classification. A number of experiments, using multiple rounds of 10-fold cross validation, are conducted on a variety of datasets to empirically evaluate the proposed approach. The results appear to be quite promising improving the performance on the classification process.