Methods of Feature Ranking and Selection for Pattern Classification

A THESIS

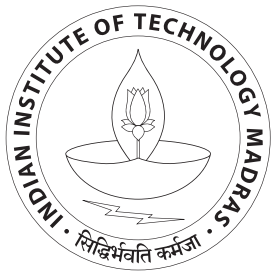
submitted by

SURANJANA SAMANTA

for the award of the degree

of

MASTER OF SCIENCE
(by Research)



DEPARTMENT OF COMPUTER SCIENCE
AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY, MADRAS.
CHENNAI-600036, INDIA.

May 2010

ABSTRACT

Pattern Classification is an important task with a wide range of applications. In case of real world texture analysis many redundant and noisy features are extracted, which degrade the performance of texture segmentation. For optimal classification performance and computational time, noisy or redundant features present in the dataset should be removed. Most of the existing feature selection methods are time consuming and parameter dependent, where the user has to specify either the total number of features in the final set or a threshold as the stopping criterion for the selection algorithm.

We propose two (unsupervised and supervised) feature ranking and selection methods, which adaptively select the optimal number of features in the dataset. In both these cases we propose a new algorithm for adaptive feature selection, where the feature ranking phase precedes feature selection. In the unsupervised case, we propose a new correlation function for feature ranking and use an existing correlation function for feature selection for unsupervised texture segmentation. The proposed method provides good performance when compared with state-of-the-art methods, when verified using benchmark real world databases and texture images.

In case of supervised feature ranking method, we propose three different criteria for feature ranking based on the following hypothesis: an ideal feature should form a set of clusters exactly equal to the number of classes in the datasets, and each of the clusters formed should be compact and non-overlapping. We determine how good

or bad a feature or a subset of features is by observing the clusters formed. After ranking, we use Fisher's criterion to select the optimum number of rank-ordered features for final task of pattern classification.

However, in case of certain class distributions, Fisher's criterion is not an ideal measure for observing class separability, and hence cannot be successfully used for feature ranking or selection. We have proposed a modified Fisher's criterion using the concept of subclass modeling. We first partition the data for each class into subclasses using a Gaussian Mixture Model. Next, we calculate Fisher's criteria on the partitioned dataset, to be used for feature ranking. The performance of both of these proposed supervised methods have been compared with state-of-the-art methods, using benchmark real world datasets.

The last part of the work involves the design of a framework, for improving the classification accuracy, by combining both decision fusion and feature selection strategies. At first, the undesired features are removed using a supervised feature selection method (classifier dependent) and each feature subset is given to the respective classifiers. Decision fusion of the classifiers provides improved classification results. For practical applications, results are shown using datasets from UCI repository and VisTex texture databases.

KEYWORDS: Feature ranking, Feature selection, Feature fusion, Dimensionality reduction, Correlation, Fisher's criterion, Subclass modeling, Supervised classification.