International Journal of Innovative
Computing, Information and Control
Volume **7**, Number **2**, February **2011**

ICIC International ©2011 ISSN 1349-4198
pp. **525–536**

EVALUATION OF DISCRIMINATION POWER OF FEATURES IN THE PATTERN CLASSIFICATION PROBLEM USING ARIF INDEX AND ITS APPLICATION TO PHYSIOLOGICAL DATASETS

Muhammad Arif

Department of Computer Science and Engineering
Air University
PAF Complex, E-9, Islamabad, Pakistan
arif@mail.au.edu.pk

Received August 2009; revised January 2010

Abstract. *In this paper, a novel index called Arif Index is proposed to evaluate the discrimination power of the features in pattern classification. Optimizing the performance of a classifier requires a prior knowledge of maximum achievable accuracy in the pattern classification using a particular set of features. Moreover, it is also desirable to know that this set of features is separable by a decision boundary of any arbitrary complexity or not. Proposed index is model free and requires no clustering algorithm to discover the clustering structure present in the feature space. It is only based on the information of local neighborhood of feature vectors in the feature space. This index can be used to predict the classification accuracy and density of the feature vectors of a class in the feature space. It was found in this paper that predicted accuracy and Arif index are very strongly correlated with each other ($R^2 = 0.99$ with p-value nearly equals to zero). This index is designed to predict the maximum achievable accuracy by a particular set of features. Implementation of the index is simple and time efficient. Performance of Arif index on different benchmark physiological data sets is found to be in consistent with the reported accuracies in the literature. Hence, this index will be very useful in providing prior useful information about the quality of features before designing any classifier.*

Keywords: Clustering, Pattern classification, Features, Nearest neighbor search, Classification accuracy

1. Introduction.

In a pattern classification problem, classification accuracy depends on a proper selection of features that can discriminate different classes and a good design of a classifier. Good design of a classifier means an ability of the classifier to approximate decision boundary of an arbitrary complexity among different classes in the feature space. Discrimination power of features decides maximum possible classification accuracy achievable by a perfect classifier. A prior knowledge of the maximum achievable classification accuracy can help a lot in deciding appropriate optimal classifier with supervised learning. Moreover, it is possible to achieve maximum classification accuracy if classes are separable in the feature space by a decision boundary of any arbitrary complexity. In real life applications [1-4], multiple clusters scattered in the feature space can represent a class. These clusters can be point, line or any arbitrary shaped clusters. Moreover, they can be compact and well separated (separated by large margin) within class or overlapping each other. Overlapping of intra-class clusters has less effect on the classification accuracy as compared to the overlap among inter-class clusters. Hence, for a good classification accuracy, clusters in a particular class need not to be very compact and well separated in the feature space but the decision boundary among classes should be well separated and