An Empirical Comparison of Hierarchical vs. Two-Level Approaches to Multiclass Problems

Suju Rajan and Joydeep Ghosh

Laboratory of Artificial Neural Systems
Department of Electrical and Computer Engineering
The University of Texas at Austin, Austin TX-78712, USA
{rsuju,ghosh}@lans.ece.utexas.edu

Abstract. The Error Correcting Output Codes (ECOC) framework provides a powerful and popular method for solving multiclass problems using a multitude of binary classifiers. We had recently introduced [10] the Binary Hierarchical Classifier (BHC) architecture that addresses multiclass classification problems using a set of binary classifiers organized in the form of a hierarchy. Unlike ECOCs, the BHC groups classes according to their natural affinities in order to make each binary problem easier. However, it cannot exploit the powerful error correcting properties of an ECOC ensemble, which can provide good results even when the individual classifiers are weak. In this paper, we provide an empirical comparison of these two approaches on a variety of datasets, using well-tuned SVMs as the base classifiers. The results show that while there is no clear advantage to either technique in terms of classification accuracy, BHCs typically achieve this performance using fewer classifiers, and have the added advantage of automatically generating a hierarchy of classes. Such hierarchies often provide a valuable tool for extracting domain knowledge, and achieve better results when coarser granularity of the output space is acceptable.

1 Introduction

Classification techniques such as the k-nearest neighbors and multi-layered perceptron can directly deal with multiclass problems. However, in difficult pattern recognition problems involving a large number of classes, it has often been observed that obtaining a classifier that discriminates between two (groups of) classes is much easier than one that simultaneously distinguishes among all classes. This observation has prompted substantial research on using a collection of binary classifiers to address multiclass problems. Further interest in this area has been fuelled by the popularity of classifiers such as SVMs [1] whose native formulation is for binary classification problems. While several extensions to multiclass SVMs have been proposed, a careful study in [2] showed that none of these approaches are superior to using a set of binary SVMs in an “All Pairs” framework.

The One-Vs-All method [3], the Round Robin Method [4] or the All Pairs method [5] (also known as pairwise method), and the Error Correcting approaches [6] [7] [8] are some of the techniques proposed for solving multiclass problem by decomposing the output space. All of these methods can be unified under a common framework wherein the output space is represented by a binary code matrix, which depends on
the technique being used [7] [9]. The above techniques can also be considered as two-level approaches in which the second level provides some relatively simple mechanism to output the final class label based on the decisions obtained from a set of binary “base” classifiers at the first level. Another common characteristic of these methods is that they do not take into consideration the affinities among the individual classes. As a result, some of the groupings might result in complex decision boundaries. In particular, this suggests the need to use powerful base classifiers in the ECOC framework (decision stumps will not do!). The performance of ECOC methods also hinges heavily on the choice of the code matrix, and though multiple solutions have been proposed, none of them are guaranteed to produce the optimal matrix for a problem at hand.

The BHC [10] is a multiclassifier system that was primarily developed to deal with multiclass hyperspectral data. Like the multiclassifier systems mentioned above, the BHC decomposes a \(C\) class problem into \((C-1)\) binary meta-class problems. However, the grouping of the individual classes into meta-classes is determined by the class distributions. A recent work [11], that compared the performance of BHC against that of an ECOC system, while using “base” Bayes classifiers in both systems, showed the superior performance of the BHC method on real-world hyperspectral data. Since the BHC not only yields valuable domain knowledge, but also resolves the problem of having to come up with an optimal code matrix, we decided to evaluate the performance of the BHC against the ECOC methods on a wide range of standard datasets. SVM, being a popular and powerful binary classifier was used as the “base” classifier in both the systems. Our experiments show that the performance of the BHC is comparable to that of ECOC classifiers while remaining robust for small training sets. We also show that besides using far lesser number of classifiers in most cases, the binary trees produced by the BHC are consistent with those a human expert might have constructed when given just the class labels.

2 Background

2.1 Binary Hierarchical Classifier

The Binary Hierarchical Classifier (BHC) [10] involves recursively decomposing a multiclass (C-classes) problem into (C-1) two meta-class problems, resulting in (C-1) classifiers arranged as a binary tree. The given set of classes is first partitioned into two disjoint meta-classes and each meta-class thus obtained is partitioned recursively until it contains only one of the original classes. The number of leaf nodes in the tree is thus equal to the number of classes in the output space. The partitioning of a parent set of classes into two meta-classes is not arbitrary, but is obtained through a deterministic annealing process, which encourages similar classes to remain in the same partition [12]. Thus, as a direct consequence of the BHC algorithm, classes that are similar to each other in the input feature space are lumped into the same meta-class higher up in the tree. Interested readers are referred to [10] for details of the algorithm. Also, note that the BHC is an example of a coarse-to-fine strategy, which has seen several application-specific successes and for which solid theoretical underpinnings are beginning to emerge [13].