Information Sciences 349–350 (2016) 102–118

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A Hierarchical Ensemble of ECOC for cancer classification based on multi-class microarray data

Kun-Hong Liu[a,b,1,*], Zhi-Hao Zeng[a], Vincent To Yee Ng[b,1]

[a] School of Software, Xiamen University, Xiamen, Fujian Province, China
[b] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

ARTICLE INFO

Article history:
Received 17 February 2015
Revised 30 September 2015
Accepted 13 February 2016
Available online 22 February 2016

Keywords:
Error Correcting Output Codes (ECOC)
Ensemble learning
Cancer classification
Feature selection
Multi-class microarray data

ABSTRACT

The difficulty of the cancer classification using multi-class microarray datasets lies in that there are only a few samples in each class. To effectively solve such a problem, we propose a hierarchical ensemble strategy, named as Hierarchical Ensemble of Error Correcting Output Codes (HE-ECOC). In this strategy, different feature subsets extracted from a dataset are used as inputs for three data-dependent ECOC algorithms, so as to produce different ECOC coding matrices. The mutual diversity degrees among these coding matrices are then calculated based on two schemes, named as the maximizing local diversity (MLD) and the maximizing global diversity (MGD) schemes. Both schemes can choose diverse coding matrices generated by the same or different ECOC algorithm(s), and the average fusion scheme is used to fuse the outputs of base learners. In the experiments, it is found that both MLD and MGD based HE-ECOC strategies work stably, and outperform individual single ECOC algorithms. In contrast with some ensemble systems, HE-ECOC generates a more robust ensemble system, and achieves better performance in most case. In short, HE-ECOC is a promising solution for the multi-class problem. The matlab code is available upon request.

1. Introduction

1.1. Background

With the development of microarray technology, it is possible to diagnose and classify some particular cancers directly based on DNA microarray datasets. Up to now, more and more new prediction, classification and clustering techniques are being used for the analysis of the data. For example, Golub et al. utilized a nearest-neighbor classifier method for the classification of acute myeloid lymphoma (AML) and acute leukemia lymphoma (ALL) in children [22]. And some studies have been reported on the application of microarray gene expression data analysis for molecular classification of cancer [2,46]. In short, microarray experiments lead to a more complete understanding of the molecular variations among tumors, and hence achieve finer and more reliable classification. Although it provides a gold mine of biological information and knowledge, it brings new challenges for biologists, statisticians and machine learning researchers because the number of tumor samples collected tends to be much smaller than the number of genes. That is, the number for the former tends to

* Corresponding author at: School of Software, Xiamen University, Xiamen, Fujian Province, China. Tel.: 86 0592 2580600.
E-mail address: lkhqz@xmu.edu.cn, lkhqz@163.com (K.-H. Liu).
[1] The first and the third authors contributes equally.

http://dx.doi.org/10.1016/j.ins.2016.02.028
0020-0255/

be on the order of tens or hundreds, while microarray datasets typically contain thousands of genes on each chip. As it is a typical "large p, small n" problem [47], an efficient and effective method for the gene expression data analysis is still a challenge.

In the field of machine learning and pattern recognition, the goal of a classification algorithm is to search a mapping function: f: $S \rightarrow K$, in which S is a set of samples and K is the corresponding class labels. For a binary class problem, there are already many widely used machine learning algorithms for estimating f. However, for a multi-class problem with N classes, a single learner is hard to produce accurate results in most cases. What's more, many elaborate classifiers, such as support vector machine (SVM), can only deal with binary problems by nature. In order to solve a multi-class problem, an alternative is to use the divide and conquer method to divide the original multi-class problem into multiple binary class problems. After dealing with each binary classification problem independently, a multi-class classification task can be handled by combining their outputs with some fusing schemes, such as the majority voting scheme. In this way, the multiclass problem is tackled by a multiple classifier system (MCS), which is proved to be more accurate and robust than an excellent single classifier in many fields [31].

With this idea, there are three basic approaches: flat strategy, hierarchical strategy and Error-Correcting Output Codes (ECOC). For the flat strategy, a fixed decomposition method, such as One vs. One (OVO) or One vs. Rest (OVR), is used, and the final label is decided directly by voting. The hierarchical strategy employs a binary tree to represent a class decomposition scheme. In the tree, each branch node represents a binary classifier and each leaf node represents a class. As for ECOC, it works in two steps: the encoding step and the decoding step. In the encoding step, the original multi-class problem is divided into multiple binary problems, represented by an $N \times M$ coding matrix: N rows represent N classes, and M columns represent a decomposition scheme using M dichotomizers. In the decoding step, the outputs of all dichotomizers are compared with each row in the coding matrix, and only the row achieving the minimum distance is selected as the final decision [17]. Hence ECOC can be considered as a more general solution than the flat and hierarchical strategies because all decomposition schemes of the former two strategies can be regarded as some special ECOC coding matrices. In addition, Kong and Dietterich proved that ECOC can reduce bias and variance errors produced by the binary classifiers more effectively [28]. So in this study, our work focuses on the application of ECOC to the field of microarray analysis.

1.2. Related work

The ECOC framework has been successfully applied in a wide range of research fields. In different studies, researchers found that it is a hard job to produce the optimum coding matrix, especially when N is large. So in recent years, many researchers devoted to the study of optimal ECOC algorithms from different perspectives. Different encoding and decoding schemes have been extensively studied. Masulli and Valentini [36] analyzed different factors that affect the effectiveness of ECOC algorithms, and found that the effectiveness of ECOC depends on the correlation between each code-word pair, structure and accuracy of dichotomizers, along with data complexity. Up to now, there is still no universal rule to design an optimal ECOC coding matrix method, and Crammer and Singer [10] proved that the design of ECOC associated to the problem domain is an NP-complete problem. So many researchers tried to use heuristic algorithms in the coding process to obtain compact coding matrix with fewer dichotomizers. It was discussed that the optimum number of dichotomizers is $[10\log_2 N, 15\log_2 N]$ [1]. Bautista et al. [5] proposed minimal ECOC to optimize ECOC coding matrix using a standard genetic algorithm (GA), and they reduced the number of binary classifiers to $[\log_2 N]$ without losing discriminative ability. Garcia-Pedrajas and Fyfe [21] used the CHC based GA to optimize the sparse random coding matrix, the length of which is limited within [30,50] regardless of the distribution of training data. Lorena and Carvalho [34] combined GA with the sparse random coding matrix too, and limited the length of code within $[\log_2 N, N]$. Furthermore, Bautista and Escalera proposed a new genetic operator to avoid invalid individuals and reduced the search space of the GA [6]. However, these methods are all based on data independent coding methods. It should be noted that the predefined coding schemes, such as OVO, random-based and all aforementioned ECOC schemes, cannot promise the best solutions for classification tasks: all such schemes neglect the data distributions.

To obtain better solutions, many researchers designed different data dependent methods by taking data distributions into consideration. For example, in [4], different feature subsets are applied to train different dichotomizers, so as to make them more independent. In [16], the original data are divided into different subclasses based on the maximum mutual information between the data and their respective class labels. Discriminated ECOC (DECOC) [38] method can generate a hierarchical partition of the class space to maximize the difference among classes by using $N-1$ dichotomizers. The partitions of a problem are learned by means of a binary tree structure using exhaustive search or a SFFS criterion, and each internal node of the tree is embedded as a column in the coding matrix. Forest-ECOC [15] uses $(N-1) \times T$ dichotomizers, where T stands for the number of binary tree structures to be embedded. It uses the classification score to create each node for building trees, and finds the trees with the maximum scores at each node to construct a robust ensemble system. ECOC-ONE [14] uses a validation set to extend the initial matrix, and increases its generalization ability by including new dichotomizers focusing on the hard classes. And it requires $2 \times N$ dichotomizers.

Although ECOC algorithms have already been well studied, its application to microarray data is just at the beginning. The difficulty of microarray data classification problem lies in the small sample size problem, which makes a validation set unaffordable. In recent years, many methods were proposed to deal with multiclass microarray datasets. For example, a gene pairs based SVM ensemble system was designed to realize the oncogene recognition and cancer classification simultaneously