Artificial Intelligence in Medicine xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

A novel hierarchical selective ensemble classifier with bioinformatics application

Leyi Wei [a], Shixiang Wan [a], Jiasheng Guo [b], Kelvin KL Wong [c,*]

[a] School of Computer Science and Technology, Tianjin University, Tianjin, China
[b] School of Information Science and Technology, Xiamen University, Xiamen, China
[c] School of Medicine, Western Sydney University, Sydney, Australia

ARTICLE INFO

Article history:
Received 23 December 2016
Received in revised form 9 February 2017
Accepted 10 February 2017

Keywords:
Selective ensemble learning
Parallel optimization
Divide and conquer
Multi-class classification
Bioinformatics

ABSTRACT

Selective ensemble learning is a technique that selects a subset of diverse and accurate basic models in order to generate stronger generalization ability. In this paper, we proposed a novel learning algorithm that is based on parallel optimization and hierarchical selection (PTHS). Our novel feature selection method is based on maximize the sum of relevance and distance (MSRD) for solving the problem of high dimensionality. Specifically, we have a PTHS algorithm that employs parallel optimization and candidate model pruning based on k-means and a hierarchical selection framework. We combine the prediction result of each basic model by majority voting, which employs the divide-and-conquer strategy to save computing time. In addition, the PT algorithm is capable to transform a multi-class problem into a binary classification problem, and thereby allowing our ensemble model to address multi-class problems. Empirical study shows that MSRD is efficient in solving the high dimensionality problem, and PTHS exhibits better performance than the other existing classification algorithms. Most importantly, our classifier achieved high-level performance on several bioinformatics problems (e.g. tRNA identification, and protein-protein interaction prediction, etc.), demonstrating efficiency and robustness.

1. Introduction

Generalization performance is one of the focuses of machine learning. An empirical study has demonstrated that ensemble learning methods offer significant improvements in generalization ability [1]. In recent years, ensemble learning methods have become a hot topic and have gained widespread applications in various fields, including multi-label classification, bioinformatics, and image processing [2–6].

In general, an ensemble learning model is an ensemble of basic models (e.g. decision trees, neural networks, and support vector machines [7–9]) that makes predictions by fusing the prediction results of basic models with some strategies, such as weighted or unweighted voting [1]. The construction process of an ensemble method can be roughly divided into two main steps. The first step is the generation and selection of the basic models. The second step is the combination of individual basic models. Bagging and Boosting are currently two well-known ensemble learning algorithms. The bagging algorithm, which is originally proposed by Breiman et al. [10], constructs a set of diverse models by the Bootstrap method that selects the samples from the original data by repeatedly back-sampling. The boosting algorithm is a type of ensemble method. AdaBoost is one of the most popular algorithms in Boosting [11]. AdaBoost is an iterative algorithm that minimizes the weighted error of the training set when training basic models. AdaBoost can increase the error classification instance weights and reduce the weight of the correctly classified instances, which leads the model to focus more on the error classification in the next iteration.

Based on intuition and experience, a growing number of basic models can improve the prediction performance. However, this approach requires more time and computing power. And, it probably even makes the performance decline if bad basic models were involved. Diversity of basic models is one of the most fundamental factors that impact the generalization performance of an ensemble model [12]. How to construct a set of diverse basic models is an important issue. Zhou et al. proposed the concept of selective ensemble and proved that it was better to create an ensemble of many basic models than to use all of the basic models [12]. They proposed an algorithm named genetic algorithm-based selective ensemble (GASEN), which selects appropriate models for ensembles. The results showed that this method could obtain good

* Corresponding author at: School of Medicine, Western Sydney University, Locked Bag 1797, Penrith, NSW, 2751, Australia.
E-mail address: Kelvin.Wong@westernsydney.edu.au (K.K. Wong).

http://dx.doi.org/10.1016/j.artmed.2017.02.005
0933-3657/

performance in smaller size models but stronger generalization performance. It is essential to choose a set of basic models that have the largest diversity instead of an ensemble of the overall basic models.

Selective ensemble learning is a method that identifies the optimal subset for creating an ensemble from a given pool of basic models by using a form of evaluation [13]. This problem is a combinatorial optimization problem that is recognized as Non-deterministic Polynomial-complete (NP). It is not easy to select the optimal subset of combined models. People have usually employed heuristic algorithms to find an approximate optimal solution, including a simulated annealing algorithm, genetic algorithm, and particle swarm optimization algorithm [13]. Kosztyán et al. proposed a novel algorithm which an optimal resource allocation with minimal total cost for any arbitrary project can be determined [14]. R. Mátrai et al. tested the characteristic searching routes and navigation methods on a web-page difference between normal users and those with intellectual disabilities [15]. Dosa et al. proved that there does not exist any polynomial-time algorithm with worst-case ratio smaller than 2 unless P = NP, even if all jobs have unit processing time [16].

In the real world, most of a practical classification task is to learn an appropriate function that assigns given input feature values into one of a finite set of classes [17]. The features are important for training a well-performing model. The upcoming problem is: should we obtain more features to build the model? Our answer: absolutely not. In fact, many features are redundant, which means that they do not affect the target class in any way. Only some of the features make the decision. Additionally, higher-dimensional features also create higher computational costs, and the performance of a model can sometimes decrease when a feature set includes irrelevant, redundant or (some) bad features. In many fields, the algorithm model may not work if the size of the dataset is too large [18]. Choosing a subset of irrelevant/redundant features is very necessary. A number of methods of ranking features have been proposed and some of them find the optimal subset of features based on cost functions. They might use exhaustive algorithms to search the best feature subset among the competing 2^N candidate subsets of features when the feature number is N. This method can find the best subset in theory, but it costs too much time. Other methods use heuristic algorithms to search for a compromising subset with a stopping condition to avoid exhaustive searching.

Multi-class classification problems have attracted a lot of attention in the machine learning field. The most typical application of multi-class classification is text categorization. Most of traditional machine learning algorithms are unable to address multi-class classification problems, because they are more complex than traditional binary classification. In a binary classification problem, each labelled sample belongs to one certain target class; while in a multi-class classification problem, each labelled sample belongs to one or more classes simultaneously [19]. Two methods were employed to address multi-class problems, including algorithm adaptation (AA) and problem transformation (PT) [20], which addresses the multi-class problem by expanding the single-label algorithm directly for multi-class classification, such as K-Nearest Neighbour, Decision trees, Naïve Bayes, and Boosting. PT transforms the multi-class problem into multiple binary problems, and then, uses traditional classification algorithms to address these problems. The PT method is more flexible and scalable than the AA method because it fully makes use of single-label algorithms for classification.

Ensemble methods can improve prediction by combining diverse basic models. However, with a large increase in the number of data sets, it becomes especially challenging to load all of the data sets into memory to train basic models. Feature selection, which aims to select the most important subset of a larger number of features, is frequently employed to address problems with large- or high-dimensional data sets [21–23]. In this paper, we propose maximize the sum of relevance and distance (MSRD), a feature selection method for solving problems with high-dimensional data sets. Drawing from ensemble theory, we also design a selective ensemble algorithm called PTHS (Parallel Optimization and Hierarchical Selection). The combination voting task of PTHS is based on a divide and conquer strategy for saving time, especially when the number of prediction instances or ensemble models is too large. We combined the two methods together to solve higher-dimensional data problems and improve the generalization performance. Finally, we used the PT method to transform the multi-class problem into a single-class problem, making our ensemble model have the ability to solve multi-class classification problems.

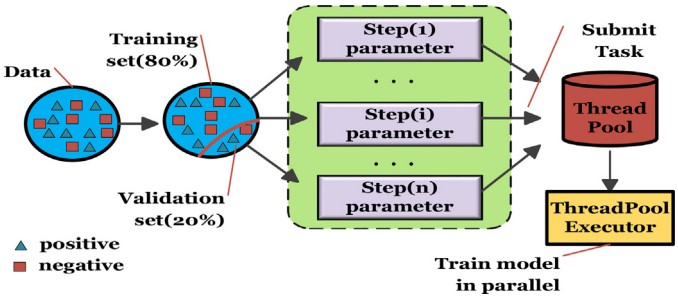

Fig. 1. Parallel optimization of a basic model.

2. Methodology design of selective ensemble

This section introduces the proposed PTHS algorithm's process in detail. The algorithm consists of three steps – parallel optimization of the basic models' parameters, ensemble pruning based on k-means clustering and hierarchical selection, and voting for the last prediction result with the divide-and-conquer strategy. Additionally, we employ the problem transformation (PT) method to expand our ensemble model for multi-class problems. Each step is described in a corresponding subsection.

2.1. Parallel optimization of the basic model's parameters

In this section, each basic model is trained independently, and the parameters are optimized in parallel. Constructing multiple basic models is the ensemble's precondition. Here, we use 22 models, $h = \{h_1, \ldots, h_i, .., h_{22}\}$, in order to generate a set of basic models that includes Logistic Regression, Random Forest, K-Nearest Neighbour (KNN), and Support Vector Machine (SVM), Naive Bayes (NB), decision trees, and other learning models [24–26]. For convenience, we consider only binary class problems here.

Let $\Omega = \{0, 1\}$ be a set of binary class labels; let $\vec{x} \in \Re^n$ be a vector that has n features that will be labelled in Ω; and let $D = \left\{\left(\vec{x}_1, l_1\right), \left(\vec{x}_2, l_2\right), \ldots, \left(\vec{x}_n, l_n\right)\right\}$ be a training set where $l_i \in \Omega$. After random sorting of the training set D, we choose 80% of D as training set and 20% as validation set. Construction of the basic models occurs in two stages. Firstly, we build the basic structure of each basic model by training the sets horizontally. When training the basic model, we optimize these models' parameters with the validation set in parallel. Because of the limited resources of the computer, we save each optimal-parameter basic model to the hard disk when it finishes parameter optimization. Note that Fig. 1 is a flowchart of the basic model's parameter optimization method.

Each step is described in details as follows:

Step 1: Perform a random sort on set, and split this set into a training set and validation set.

Step 2: Build the base structure of model h_i using the training set. Here, h_i is implemented using a data mining tool, Waikato Envi-

Please cite this article in press as: Wei L, et al. A novel hierarchical selective ensemble classifier with bioinformatics application. Artif Intell Med (2017), http://dx.doi.org/10.1016/j.artmed.2017.02.005