Applied Soft Computing 11 (2011) 5674–5683

Contents lists available at ScienceDirect

Applied Soft Computing

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

A novel ensemble algorithm for biomedical classification based on Ant Colony Optimization

Lei Shi[a,b,c], Lei Xi[c], Xinming Ma[a,b,c,*], Mei Weng[c], Xiaohong Hu[c]

[a] Agronomy College, HeNan Agricultural University, Zhengzhou 450002, China
[b] The Incubation Base of National Key Laboratory for Physiological Ecology and Genetic Improvement of Food Crops in Henan Province, HeNan Agricultural University, Zhengzhou 450002, China
[c] College of Information and Management Science, HeNan Agricultural University, Zhengzhou 450002, China

ARTICLE INFO

Article history:
Received 5 December 2010
Received in revised form 17 March 2011
Accepted 23 March 2011
Available online 2 April 2011

Keywords:
Ant Colony Optimization
Rough set
Ensemble learning
Biomedical classification

ABSTRACT

One of the major tasks in biomedicine is the classification and prediction of biomedical data. Ensemble learning is an effective method to significantly improve the generalization ability of classification and thus have obtained more and more attentions in the biomedicine community. However, most existing techniques in ensemble learning employ all the trained component classifiers to constitute ensembles, which are sometimes unnecessarily large and can lead to extra memory costs and computational times. For improving the generalization ability and efficiency of ensemble for biomedical classification, an Ant Colony Optimization and rough set based ensemble approach is proposed in this paper. Ant Colony Optimization and rough set theory are incorporated to select a subset of all the trained component classifiers for aggregation. Experiment results show that compared with existing methods, it not only decreases the size of ensemble, but also obtains higher prediction performance.

1. Introduction

One of the major tasks in biomedicine is the classification and prediction of biomedical data. It could lead us to the elucidation of the secrets of life or ways to prevent certain currently non-curable diseases such as HIV. Although laboratory experiment is the most effective method for investigating the data, it is very financially and labor expensive. With the rapid increase in size of the biomedical databases, it is essential to use computational algorithms and tools to automate the classification process. Now, many algorithms in the fields of machine learning have therefore been widely used for the classification analysis of biomedical data, such as decision trees, k-nearest neighbor and artificial neural network [1].

Ensemble method is one of the major advances in machine learning in the past years. It is learning algorithm that trains a set of component classifiers and then combines their predictions to classify new examples [2]. As an effective method to improve classification performance, ensemble technique is available for the classification analysis of biomedical data and thus gaining more and more attentions in biomedicine community. However, ensemble method has two important drawbacks. Firstly, it requires much more memory to store all the learning models in the ensemble, and secondly it takes much more computation time to produce a prediction for an unlabeled example. The storage and computation time increase with the number of component classifiers in the ensemble. Most existing techniques in ensemble learning employ all the trained component classifiers to constitute ensembles, which are sometimes unnecessarily large and can lead to extra memory costs and computational times. The problems frequently limit the application of ensemble method to classification of biomedical data.

Rough set theory, introduced by Pawlak in 1982, is a formal mathematical tool to deal with imprecision, uncertainty and vagueness [3]. As an important feature selection method, rough set can preserve the meaning of the features. The essence of rough set approach to feature selection is to find a subset of the original features. However, the number of possible subsets is always very large when N (N is the number of features) is large because there are 2^N subsets and to examine exhaustively all subsets of features for selecting the optimal one is an NP-hard problem. Therefore, it is necessary to investigate fast and effective approximate algorithms. Previous methods employed an incremental hill-climbing algorithm to select feature. However, this often led to a non-minimal feature combination.

Ant Colony Optimization (ACO) is a population-based paradigm that can be used to find approximate solutions to difficult optimization problems. The first ACO algorithm which can be classified within this technique was introduced in the early 1990s by Col-

* Corresponding author at: Agronomy College, HeNan Agricultural University, Zhengzhou 450002, China.
E-mail address: xinmingma@126.com (X. Ma).

1568-4946/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2011.03.025

orni and Dorigo [4], and since then many diverse variants of the basic principle have been reported in the literature. ACO algorithm is inspired by the social behavior of ant colonies in their search for the shortest path to food sources. Although they have no sight, ants are capable of finding the shortest route between a food source and their nest by chemical materials called pheromone that they leave when moving. As an important branch of newly developed form of artificial intelligence called Swarm Intelligence, ACO algorithm has been shown to be an effective tool in finding good solutions. It has an advantage over simulated annealing and Genetic Algorithm approaches when the graph may change dynamically because it can be run continuously and adapt to changes in real time [5]. ACO algorithm was firstly used in solving traveling salesman problem (TSP) [6] and then has been successfully applied to a large number of difficult problems like the quadratic assignment problem (QAP), routing in telecommunication networks, graph coloring problems, feature selection, etc. [7]. Particularly, ACO is attractive for feature selection since there is no heuristic information that can guide search to the optimal minimal subset every time and ants can discover the best feature combinations as they traverse the graph when features are represented as a graph.

For improving the prediction ability and efficiency to classify biomedical data, an ACO and rough set based ensemble algorithm is proposed in this paper. ACO and rough set theory are incorporated to select a subset of the all trained component classifiers for aggregation. Experiment results show that compared with existing methods, it not only decreases the size of ensemble, but also obtains higher performance of prediction for biomedical data.

The remainder of the paper is organized as follows. Section 2 gives an overview of related work. Section 3 introduces the basic background ideas about ensemble learning, rough set and ACO for the sake of further discussion. Section 4 introduces the incorporation of ACO with rough set for feature selection. Section 5 describes the proposed novel ensemble algorithm in detail. Section 6 discusses experimental results. Finally, Section 7 presents concluding remarks and directions of our future work.

2. Related work

Machine learning is the subfield of artificial intelligence which focuses on methods to construct computer programs that learn from experience with respect to some class of tasks and a performance measure [8]. Machine learning methods are suitable for biomedical data due to the learning algorithm's ability to construct classifiers that can explain complex relationships in the data. Recently, the use of machine learning has become widely accepted in biomedical applications and many researches have been conducted to address the classification of biomedical data in the literature.

In [9], the shrunken centroid method is proposed to classify biomedical data. It relies on a nearest-class centroid classification, but using the centroids of the classes shrunken towards the centroid of all classes by a threshold-controlled amount. The threshold can be determined by a cross validation process. The method is similar to linear discriminant analysis, but assumes a diagonal pooled covariance matrix. In [10], the classification methods Fisher Linear Discriminant Analysis and Least Squares SVM (linear and radial kernel) are studied for classification of biomedical data. The performance of the two methods was compared by leave-one-out cross validation and the experimental results indicate that the importance of regularizing the classifiers and suggest that the LS-SVM with the RBF kernel is prone to overfitting in biomedical classification.

Artificial neural networks have been identified as effective approach in biomedical classification. In [11], artificial neural network is used to develop a method of classifying cancers to specific diagnostic categories. The experiment demonstrates the potential applications of artificial neural network for tumor diagnosis and the identification of candidate targets for therapy.

Ensemble learning has increasingly gained attention in biomedical research. In [12], a comparison of single supervised machine learning and ensemble methods is performed in classifying seven publicly available cancerous data. The experimental results indicate that ensemble methods consistently perform well over all the datasets in terms of their specificity. A combinational feature selection and ensemble neural network method is introduced for classification of biomedical data in [13]. However, those researches of ensemble learning employ all the trained component classifiers to constitute ensembles, which are sometimes unnecessarily large and can lead to extra computational times. In some scenarios of biomedical data classification such as dynamically mining large repositories, ensemble learning is not suitable due to its lower efficiency.

Intelligent algorithms such as Genetic Algorithms (GA) [14] and Particle Swarm Optimization (PSO) [15] have been conducted to design ensembles in a number of studies. In [16], GA is employed for selecting the features as well as selecting the types of individual classifiers to design the fusion strategy of classifier. In [17], GA is used to optimize the weights of the feature vector that represents the importance of the features to obtain better prediction accuracy. In [18], a PSO based ensemble classifier is proposed and evaluated. Each nearest prototype classifier of the ensemble is generated sequentially using PSO. The PSO is used to find the prototypes' locations with the objective of reducing the error rate and the diversity among the members of the ensemble is enforced through different initialization of PSO. Simulation experiments on different classification problems show that the PSO based ensemble classifier has better performance than a single classifier. Like GA and PSO, ACO is a new evolutionary computation technique and has been applied to many combinatorial optimization problems. ACO has an advantage over PSO and GA approaches of similar problems when the graph may change dynamically and the ant colony algorithm can be run continuously and adapt to changes in real time [5]. Compared with GA and PSO, ACO requires only primitive and simple mathematical operators, and does not need complex operators such as crossover and mutation. Then, it is inexpensive in terms of both memory costs and computational times. Furthermore, ACO is particularly attractive for feature selection as there seems to be no heuristic that can guide search to the optimal minimal subset every time. Additionally, it can be the case that ants discover the best feature combinations as they proceed throughout the search space. In this paper, ACO is adopted to combine with rough set theory to select a subset of all the trained component classifiers for aggregation, and a novel improved ensemble algorithm based on ACO and rough set is proposed. Experiments are carried out on several public biomedical datasets. The experimental results indicate that the proposed approach achieves significant performance improvement.

3. Preliminaries

3.1. Ensemble learning

Ensemble learning is a method that trains a set of individual classifiers and then combines their predictions in some way to clas-