Computers in Biology and Medicine 87 (2017) 8–21

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Parallel gene selection and dynamic ensemble pruning based on Affinity Propagation

Jun Meng [a], Jing Zhang [a], Yu-Shi Luan [b,*], Xin-Yu He [a], Li-Shuang Li [a], Yuan-Feng Zhu [c]

[a] School of Computer Science and Technology, Dalian University of Technology, Dalian 116023, China
[b] School of Life Science and Biotechnology, Dalian University of Technology, Dalian 116023, China
[c] BorderX Lab Inc, Silicon Valley, California, 94086, USA

ARTICLE INFO

Keywords:
Intersection neighborhood rough set
MapReduce
Affinity Propagation
Dynamic ensemble pruning
Microarray data

ABSTRACT

Gene selection and sample classification based on gene expression data are important research areas in bioinformatics. Selecting important genes closely related to classification is a challenging task due to high dimensionality and small sample size of microarray data. Extended rough set based on neighborhood has been successfully applied to gene selection, as it can select attributes without redundancy and deal with numerical attributes directly. However, the computation of approximations in rough set is extremely time consuming. In this paper, in order to accelerate the process of gene selection, a parallel computation method is proposed to calculate approximations of intersection neighborhood rough set. Furthermore, a novel dynamic ensemble pruning approach based on Affinity Propagation clustering and dynamic pruning framework is proposed to reduce memory usage and computational cost. Experimental results on three *Arabidopsis thaliana* biotic and abiotic stress response datasets demonstrate that the proposed method can obtain better classification performance than ensemble method with gene pre-selection.

1. Introduction

Plants often encounter various stresses at different growth stages throughout their life, which may lead to inhibition of growth, leaf injury and plant death [1]. One type of these stresses is biotic stress which is caused by the attack of other organisms, such as virus and pathogens, or herbivorous insects and parasitic plants. The other type is called abiotic stresses which is related to nonliving factors in the environment like salt, drought, heavy metals and high intensity light.

To eliminate negative influences due to these stresses before the appearance of some symptoms, gene expression data is used to diagnose and recognize the types of plant stresses. Since gene selection is a crucial step towards effective classification [2,3] based on large scale of gene data, high performance methods for gene selection and sample classification have become increasingly important.

Rough set is an important mathematical tool to deal with uncertainty and vagueness of decision system and has been successfully applied in data mining and machine learning [4–7]. Gene selection method based on rough set has demonstrated that it can consider both individual gene information and mutual information among them, and selects the gene subset without redundancy [8]. Therefore, rough set has been used to select important genes for the classification of microarray data [9]. However, since the classical rough set model proposed by Pawlak [10] only deals with data with nominal attributes, thus, numerical attributes needed to be discretized, which leads to loss of information. To solve this problem, extended rough set models have been proposed. Hu et al. presented the neighborhood relation to replace the equivalence relation in classical rough set model to process numerical attributes [11–13]. Wang et al. proposed a series of gene selection methods based on neighborhood rough set theory [14–16]. For the analysis of plant stress response, Meng et al. presented a gene selection method based on intersection neighborhood rough set [9].

With the rapid growth of the amount of data and the application of large-scale parallel computing, parallel data mining technology has been widely used. In the parallelized computation of rough set, Zhang et al. proposed a parallel algorithm for accelerating the computation process of approximations in rough set [17], and presented a comparison of rough set based knowledge acquisition on different MapReduce runtime systems [18]. Qian et al. proposed a parallel rough set attribute reduction algorithm using MapReduce programming model [19].

In order to improve the classification performance, ensemble classifier is employed in the analysis of gene expression data. A large number

* Corresponding author.
E-mail address: luanyush@dlut.edu.cn (Y.-S. Luan).

http://dx.doi.org/10.1016/j.compbiomed.2017.05.016
Received 13 January 2017; Received in revised form 13 May 2017; Accepted 13 May 2017

of base classifiers are generated for the same problem, and therefore a large amount of memory and considerable computational cost are needed [20]. To solve this problem, classifier pruning is admirable for ensemble model. Zhou revealed that sometimes ensemble of a proper subset (with high accuracy and diversity) of base classifiers outperforms that of all base classifiers, which is an unexpected advantage of ensemble pruning [21]. Most of the previous works focused on the improvement of ensemble pruning techniques, while little attention has been devoted to the development of the hybrid approach, which refers to combine other machine learning technique with classification. Lin et al. proposed an ensemble pruning method using k-means clustering and dynamic selection strategy [22]. Zhang et al. presented an ensemble pruning approach, which was based on spectral clustering [23]. Krawczyk employed clustering algorithm in ensemble pruning method for weighted bagging [24]. Dynamic ensemble pruning strategy is combined with classifier clusters generated by Affinity Propagation (AP) clustering. AP clustering has three advantages: (1) it is more efficient; (2) it is insensitive to the initialization and doesn't need to preset the number of clusters; (3) it generates clusters at a lower error rate than other exemplar-based clustering methods such as k-means [25,26]. AP clustering algorithm is composed of two steps [25]: (1) construct a similarity matrix, it takes similarity matrix as the input; (2) "responsibility" and "availability" messages are exchanged among data points until high-quality set of exemplars and corresponding clusters gradually emerge.

In this paper, parallelized gene selection and dynamic ensemble pruning based on Affinity Propagation (PGS-DEP-AP) is proposed. A parallel computation method of approximations in intersection neighborhood rough set for gene selection. In addition, a novel ensemble pruning approach is presented, which is based on AP clustering and the framework of dynamic pruning. The framework of PGS-DEP-AP method is shown in Fig. 1.

Step 1: Attribute reduction model based on the intersection neighborhood rough set selects important genes, and different significant measures are used as the heuristic information.

Step 2: Generating SVM classifier on each reduced gene subset. And all the base classifiers classify the validation samples on which the similarity matrix is generated according to the classification output.

Step 3: AP clustering algorithm takes the similarity matrix as input and groups all base classifiers into many clusters.

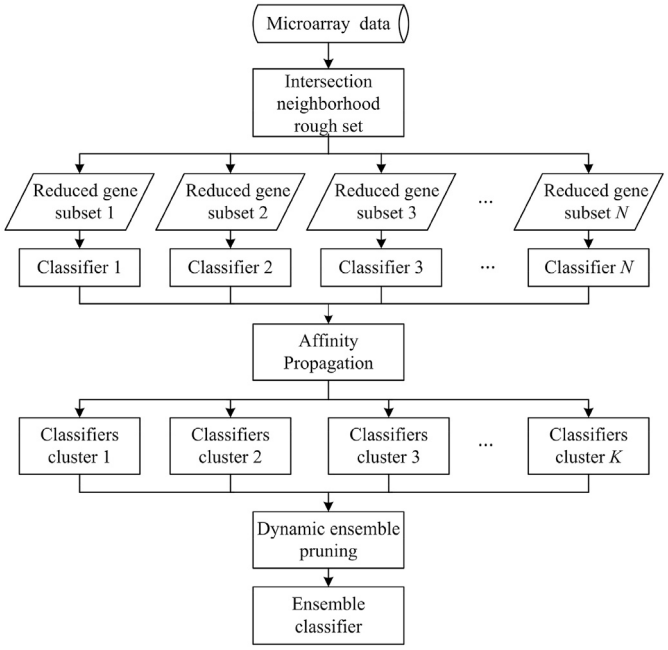

Fig. 1. Classification framework of PGS-DEP-AP.

Step 4: Dynamic ensemble pruning selects base classifiers based on classifier clusters generated by step 3.

Step 5: Each test sample is classified by all selected base classifiers, then the model integrates the results of different classifiers by majority vote method.

The rest of the paper is organized as follows: Section 2 describes the parallelized computation method of approximations in intersection neighborhood rough set, and the gene selection method based on rough set. Our proposed ensemble pruning method based AP and dynamic pruning strategy is described in Section 3. Experiment results and analysis are discussed in Section 4. Finally, conclusion and future work are given in Section 5.

2. Gene selection based on intersection neighborhood rough set

In comparison to tens of thousands of genes, gene microarray data has relatively small samples due to expensive cost of data collection. Furthermore, only a few of genes are highly associated with classification [9]. Gene selection can improve accuracy and reduce computation cost; therefore, it is very essential for the classification of microarray data.

There are generally three kinds of gene selection techniques: filter, wrapper and embedded. Filter techniques assess genes by calculating their relevance scores, and then eliminate low-scoring features [27–29]. They are independent of the classification algorithm, less computationally intensive and have faster training speed, thus, they easily scale to very high-dimensional data [30]. Wrapper techniques update the selected gene subset according to a criterion such as accuracy of the classifier [31–34]. The common drawbacks of wrapper techniques are higher risk of overfitting and intensive computation [30]. Embedded approaches also combine with a classifier, however, they have the advantage of including interaction with the classification model, while being far less computationally intensive than wrapper methods [30]. Many redundant genes may be selected by these methods, since there are only a few significant genes that are closely related to classification. Compare to the methods mentioned above, rough set based gene selection can obtain gene subsets without redundancy for classification [35–39].

The computation process of approximations and the matrix representation of intersection neighborhood rough set which is used in the parallelized computation of approximations are introduced firstly. Then, we propose a parallelized intersection neighborhood rough set computation method, which is suitable for accelerating the gene selection process using MapReduce. Finally, the gene selection model based on rough set and different heuristic information is presented to select important genes for classification.

2.1. Rough set

2.1.1. Rough set model based on intersection neighborhood

The rationale of rough set theory is that a set of distinct objects are approximated via lower and upper bound. The computation of approximations is extremely time consuming. Therefore, the sequential methods of rough set can only deal with small data sets. In order to introduce rough set into large-scale data mining, efficient computation of approximations is vital.

Gene expression data table is represented by $GEDT = \{S, G \cup D, V, f\}$, where $S = \{s_1, s_2, \ldots, s_n\}$ is the sample set; $G = \{g_1, g_2, \ldots, g_m\}$ is a set of genes called condition attributes; D is the class label called decision attribute; $V = \cup_{b \in G \cup D} V_b$ and V_b is a domain of attribute b; $f: S \times G \cup D \rightarrow V$ is the information function, for every $x \in S$, $b \in G \cup D$, $f_b(x) \in V_b$ denotes the value of sample x on the attribute b.

Definition 1. $R \subseteq S \times S$ is binary relation whereby the neighborhood of object $x \in S$ is defined as follows [40]:

$$N_R(x) = \{y | xRy, y \in S\}. \quad (1)$$