Applied Soft Computing 25 (2014) 51–63

Contents lists available at ScienceDirect

Applied Soft Computing

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

Gene selection using rough set based on neighborhood for the analysis of plant stress response

Jun Meng[a], Jing Zhang[a], Rui Li[a], Yushi Luan[b,*]

[a] School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116023, China
[b] School of Life Science and Biotechnology, Dalian University of Technology, Dalian, Liaoning 116023, China

ARTICLE INFO

Article history:
Received 10 July 2013
Received in revised form
11 September 2014
Accepted 13 September 2014
Available online 22 September 2014

Keywords:
Gene selection
Rough set based on neighborhood
Threshold optimization
Plant stress response

ABSTRACT

Gene selection and sample classification based on gene expression data are important research trends in bioinformatics. It is very difficult to select significant genes closely related to classification because of the high dimension and small sample size of gene expression data. Rough set based on neighborhood has been successfully applied to gene selection, as it selects attributes without redundancy and deals with numerical attributes directly. Construction of neighborhoods, approximation operators and attribute reduction algorithm are three key components in this gene selection approach. In this study, a novel neighborhood named intersection neighborhood for numerical data was defined. The performances of two kinds of approximation operators were compared on gene expression data. A significant gene selection algorithm, which was applied to the analysis of plant stress response, was proposed by using positive region and gene ranking, and then this algorithm with thresholds optimization for intersection neighborhood was extended. The performance of the proposed algorithm, along with a comparison with other related methods, classical algorithms and rough set methods, was analyzed. The results of experiments on four data sets showed that intersection neighborhood was more flexible to adapt to the data with various structure, and approximation operator based on elementary set was more suitable for this application than that based on element. That was to say that the proposed algorithms were effective, as they could select significant gene subsets without redundancy and achieve high classification accuracy.

1. Introduction

With the development of microarray technology and genetic sequencing technology, more and more gene expression data sets containing large amounts of information are obtained, which allowed simultaneous monitoring of the expression levels of tens of thousands of genes. A increasing number of bioinformatics researchers had paid more attention to mine gene expression data. One of the important trends in bioinformatics was the identification of genes or groups of genes to differentiate diseased tissues from normal tissues, which could be viewed as gene selection and sample classification [1]. This study was of great significance on plant stress response [2–5], tumor and cancer classification [6–9], disease diagnosis [10–12], etc.

Plants encountered numerous stresses in the field, including biotic and abiotic stresses. Biotic stresses were derived from the infection of virus and pathogens, and the invasion of herbivorous insects and parasitic plants. Others were abiotic stresses which were derived from global climate change and environmental disruption caused by industrial manufacture, such as salt, drought stress, heavy metals, high intensity light, etc. From the viewpoint of agriculture and nature conservation, it was important to diagnose the kind of plant stresses before the appearance of some symptoms, such as inhibition of growth, leaf injury and plant death [2].

It is a challenging problem to select significant genes which are closely related to classification because of the high dimension and small sample size of gene expression data. Gene selection methods can be generally divided into two categories: filter methods in which gene selection is independent of classification, and wrapper methods which combine gene selection with a classifier. Most filter methods employ gene ranking approaches in which genes are ranked according to some criterions, such as statistical test [3,6,13]. While wrapper methods update the selected gene subset iteratively, according to a criterion such as accuracy of the classifier [4,14–17]. However, these methods have some weakness. For instance, filter methods select genes only using individual contribution, while ignoring mutual information among genes. Wrapper

* Corresponding author. Tel.: +86 411 84706356; fax: +86 411 84706365.
E-mail addresses: mengjun@dlut.edu.cn (J. Meng), danna-zj@163.com (J. Zhang), lr91711@163.com (R. Li), luanyush@dlut.edu.cn (Y. Luan).

http://dx.doi.org/10.1016/j.asoc.2014.09.013
1568-4946/

methods are unstable to some extent, because the accuracy is sensitive to the selected gene subset [10]. Meanwhile, many redundant genes may be selected by these methods, since there are only several significant genes closely related to classification.

Rough set theory, proposed by Pawlak in 1982 [18], has been successfully applied to gene selection, as it can get rid of redundant attributes using individual attribute information and mutual information among them. Rough set theory is a useful tool to deal with vague, uncertain and incomplete information. A number of the researches on its theories and applications were reported [18–24]. The selection criterions were constructed, which used attribute dependence and significance measure based on classical rough set model, and employed those criterions in gene selection [25–27]. However, classical rough set model could only deal with data with nominal attributes, so the data needed to be discretized before gene selection, which led to information dropout. To address this problem, many extended rough set models were employed. Variable precision rough set (VPRS) were introduced, which allowed identifying data patterns that otherwise would be lost, to the selection and classification of gene expression data [28]. They selected the most relevant supergenes by two methods, the criterion of maximum β-relevance and the VPRS-Q algorithm, both supported by VPRS. Dai and Xu [7] proposed an attribute selection method for gene expression data based on fuzzy gain ratio under the framework of fuzzy rough set theory which combined fuzzy set theory and rough set theory. In addition, Wang et al. [8,9,29] proposed a series of gene selection and classification methods using neighborhood rough set model defined by Hu et al. [30–32]. They employed δ neighborhood to deal with numerical data directly, and used forward attribute reduction algorithm to select genes.

Three key components in gene selection methods using rough set model based on neighborhood are construction of neighborhoods, approximation operators and attribute reduction algorithm. How to construct neighborhoods to suit for various data structures, and design more effective attribute reduction algorithm, are problems to be solved. Study on those components is meaningful for the development of gene selection based on expression data. In this study, we defined a novel neighborhood for numerical data, named intersection neighborhood, which could be more flexible to adapt to data with various structure. The performances of two kinds of approximation operators on expression data were compared. We proposed a significant gene selection algorithm using positive region and gene ranking, and then extended this algorithm with thresholds optimization for intersection neighborhood. The proposed algorithms were applied to the analysis of plant stress response.

The remainder of this paper is organized as follows. In Section 2, some basic concepts about rough set and neighborhood are introduced, and intersection neighborhood is defined. The framework of gene selection using rough set model based on neighborhood is elaborated and a novel significant gene selection algorithm is proposed in Section 3. In Section 4, an extended algorithm with thresholds optimization for intersection neighborhood is proposed. Compared with other related methods, classical algorithms and rough set algorithms, experimental results are shown in Section 5. Finally, Section 6 concludes this paper.

2. Rough set model based on neighborhood

The main idea of rough set theory is building elementary sets of knowledge representation on object space, and then using two sets, which can be represented based on elementary sets, to approximate an indescribable set. Rough set model based on neighborhood is an extended model of Pawlak classical rough set model, in which neighborhoods are elementary sets [19,33]. In this section, we discuss two key definitions in this extended model in detail, neighborhood and approximation, and then propose a new definition of neighborhood.

2.1. Definitions of neighborhood

A neighborhood of an object x is a set of objects with similar characteristics to x. A generalized definition for neighborhood has been given according to binary relation [19].

Definition 2.1 *([19])*. For an object $x \in U$, U is a nonempty finite set of objects called the universe, and a binary relation R on U, the neighborhood of x is:

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

On the basis of the generalized definition, many specific definition formulas of neighborhood are proposed to deal with complex real data. As all data in gene selection are numerical attributes, we focus on the definitions of neighborhood for numerical attributes, which are usually induced by metric functions. The δ neighborhood defined by Hu et al. is an effective way to deal with numerical attributes.

Definition 2.2 *([31])*. For an object $x \in U$ and a subset of attributes $B \subseteq C$, the δ neighborhood of x induced by B is defined as:

$$\delta_B(x) = \{y | \Delta_B(x, y) \le \delta, y \in U\}, \tag{2}$$

where $\Delta_B(x,y)$ is a distance metric function to determine the shape of neighborhood and δ is a threshold to control the size of neighborhood. For numerical attributes, a general used metric, named Minkowsky distance, is defined as:

$$\Delta_B(x, y) = \left(\sum_{b \in B} \left| f_b(x) - f_b(y) \right|^p \right)^{1/p}, \tag{3}$$

where (1) it is called Manhattan distance, if $p = 1$; (2) Euclidean distance, if $p = 2$; (3) Chebychev distance, if $p = \infty$.

The δ neighborhood can be represented with the form of neighborhood defined in Definition 2.1 by constructing a binary relation based on the metric function. The binary relation is:

$$\Delta_B^\delta = \{(x, y) \in U \times U | \Delta_B(x, y) \le \delta\}. \tag{4}$$

It is shown that this binary relation is a tolerance relation as it is reflexive and symmetric but not transitive.

In the gene selection based on gene expression data, Euclidean distance is the most frequently used. The δ neighborhood mentioned in the rest of this study is all based on Euclidean distance unless otherwise specified. However, there are some limitations of δ neighborhood. On one hand, the shape of δ neighborhood is a circular for two attributes, sphere for three attributes and hypersphere for more attributes, so this neighborhood is not suited to deal with data in which the structure is square, cube or hypercube. On the other hand, the threshold δ is not changed with different attributes or different number of attributes, because it is hardly realistic to determine different value of δ for so many possible cases. In order to deal with the problems above, a new definition of neighborhood named intersection neighborhood is proposed.

Definition 2.3. For an object $x \in U$ and a subset of attributes $B \subseteq C$, the intersection neighborhood of x induced by B is defined as:

$$N_B^\delta(x) = \{y | \forall b \in B, |f_b(x) - f_b(y)| \le \delta_b, y \in U\}, \tag{5}$$

where δ_b is a specific threshold for the attribute b.