Artificial Intelligence in Medicine (2008) **44**, 7—25

ARTIFICIAL
INTELLIGENCE
IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

CARSVM: A class association rule-based classification framework and its application to gene expression data

Keivan Kianmehr [1], **Reda Alhajj** *

BIDEALS Group, Department of Computer Science, University of Calgary, 2500 University Drive NW, Calgary, Alberta, Canada T2N 1N4

Received 12 December 2007; received in revised form 10 May 2008; accepted 13 May 2008

KEYWORDS
Machine learning;
Association rule mining;
Associative classifiers;
Support vector machine;
Data mining;
Gene expression analysis;
Gene expression classification;
Gene selection

Summary

Objective: In this study, we aim at building a classification framework, namely the CARSVM model, which integrates association rule mining and support vector machine (SVM). The goal is to benefit from advantages of both, the discriminative knowledge represented by class association rules and the classification power of the SVM algorithm, to construct an efficient and accurate classifier model that improves the interpretability problem of SVM as a traditional machine learning technique and overcomes the efficiency issues of associative classification algorithms.
Method: In our proposed framework: instead of using the original training set, a set of rule-based feature vectors, which are generated based on the discriminative ability of class association rules over the training samples, are presented to the learning component of the SVM algorithm. We show that rule-based feature vectors present a high-qualified source of discrimination knowledge that can impact substantially the prediction power of SVM and associative classification techniques. They provide users with more conveniences in terms of understandability and interpretability as well.
Results: We have used four datasets from UCI ML repository to evaluate the performance of the developed system in comparison with five well-known existing classification methods. Because of the importance and popularity of gene expression analysis as real world application of the classification model, we present an extension of CARSVM combined with feature selection to be applied to gene expression data. Then, we describe how this combination will provide biologists with an efficient and understandable classifier model. The reported test results and their biological interpretation demonstrate the applicability, efficiency and effectiveness of the proposed model.

* Corresponding author. Tel.: +1 403 220 9453; fax: +1 403 284 4707.
E-mail addresses: mkkian@ucalgary.ca (K. Kianmehr), alhajj@ucalgary.ca, rsalhajj@gmail.com (R. Alhajj).
[1] Tel.: +1 403 220 9664; fax: +1 403 284 4707.

doi:10.1016/j.artmed.2008.05.002

Conclusion: From the results, it can be concluded that a considerable increase in classification accuracy can be obtained when the rule-based feature vectors are integrated in the learning process of the SVM algorithm. In the context of applicability, according to the results obtained from gene expression analysis, we can conclude that the CARSVM system can be utilized in a variety of real world applications with some adjustments.

1. Introduction

Classification is an important data mining task, widely used in numerous real world applications. It aims at exploring through data objects (training set) to find a set of rules which determine the class of each object according to its attributes. These rules are later used to build a classifier to predict the class or missing attribute value of unseen objects whose class might not be known. A wide variety of research has considered the use of popular machine learning techniques, such as neural network, decision trees, Bayesian network and support vector machines, in classification problems. Despite their good performance in real world applications, machine learning techniques have some shortcomings. They work based on mathematical and statistical algorithms and use domain independent biases to extract the rules. Therefore, they are not able to discover all the interesting and understandable rules in the analyzed dataset. The extracted rules may not satisfy domain expert's expectations and interests as well. For instance, there might be some rules which may play an important role in understanding the classification task, but not discovered by machine learning techniques.

1.1. Associative classification

To overcome the understandability problem of the classification task [1,2], association rule-based classification techniques, known as associative classification, have been recently proposed and have received great consideration. In associative classification, a classifier is built by using a subset of association rules, namely CARs [3], where the consequent of each rule is a class attribute. For predicting the class label of a given object, the classifier analyzes a rule or a set of rules matching the object. The related literature indicates that associative classifiers have better results than machine learning classification algorithms. However, they suffer from efficiency issues. First, the rule generator algorithm generates a large number of rules, and it is difficult to store the rules, retrieve the related rules, prune and sort the rules [4]. Second, it is challenging to find the best subset of rules to build the most robust and accurate classifier.

1.2. Gene expression classification

In microarray gene expression analysis, classification is applied to discriminate diseases or to predict outcomes based on gene expression patterns, and perhaps even to identify the best treatment for given genetic signature [5]. There are several general issues in microarray classification. First, the datasets are usually complex and noisy. Noise and complexity in experimental protocols strongly limit data integration. Another important issue in gene classification is the fact that currently available datasets typically contain fewer than one hundred instances, though each instance quantifies the expression levels of several thousands of genes (i.e., high dimensionality). Due to the high dimensionality and the small sample size of the experimental data, it is often possible to find a large number of classifiers that can separate the training data perfectly, but their diagnostic accuracy on unseen test samples is quite poor and different. Therefore, traditional methods for data mining cannot be effectively applied to gene expression classification [6], and there is a need for more dedicated techniques to approach this problem. As one solution, feature construction and feature selection have been applied to pre-process the gene expression data in a way to overcome the high-dimensionality problem [6,7]. Many supervised machine learning algorithms such as neural networks, Bayesian networks and support vector machine (SVM), combined with feature selection techniques, have been previously applied to microarray gene expression. In the work described in [8], we achieved outstanding accuracy in gene expression classification by combining neural networks learning and SVM-based feature selection into an integrated approach. Although these integrated techniques to gene expression classification can achieve high accuracy, many of the rules discovered by them are not valid and understandable by biologists. Recall that, in machine learning algorithms, rule generation is based on mathematical and statistical algorithms which use domain independent biases.