Expert Systems with Applications 36 (2009) 8616–8630

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A hybrid evolutionary algorithm for attribute selection in data mining

K.C. Tan [a,*], E.J. Teoh [a], Q. Yu [a,b], K.C. Goh [a]

[a] Department of Electrical and Computer Engineering, National University of Singapore, 4 Engineering Drive 3, Singapore 117576, Singapore
[b] Rochester Institute of Technology, USA

ARTICLE INFO

Keywords:
Evolutionary algorithms
Support vector machines
Data mining
Attribute selection
Pattern classification

ABSTRACT

Real life data sets are often interspersed with noise, making the subsequent data mining process difficult. The task of the classifier could be simplified by eliminating attributes that are deemed to be redundant for classification, as the retention of only pertinent attributes would reduce the size of the dataset and subsequently allow more comprehensible analysis of the extracted patterns or rules. In this article, a new hybrid approach comprising of two conventional machine learning algorithms has been proposed to carry out attribute selection. Genetic algorithms (GAs) and support vector machines (SVMs) are integrated effectively based on a wrapper approach. Specifically, the GA component searches for the best attribute set by applying the principles of an evolutionary process. The SVM then classifies the patterns in the reduced datasets, corresponding to the attribute subsets represented by the GA chromosomes. The proposed GA-SVM hybrid is subsequently validated using datasets obtained from the UCI machine learning repository. Simulation results demonstrate that the GA-SVM hybrid produces good classification accuracy and a higher level of consistency that is comparable to other established algorithms. In addition, improvements are made to the hybrid by using a correlation measure between attributes as a fitness measure to replace the weaker members in the population with newly formed chromosomes. This injects greater diversity and increases the overall fitness of the population. Similarly, the improved mechanism is also validated on the same data sets used in the first stage. The results justify the improvements in the classification accuracy and demonstrate its potential to be a good classifier for future data mining purposes.

1. Introduction

In today's context, data mining has developed into an important application due to the abundance of data and the imperative to extract useful information from raw data. Many useful data patterns can be selected out, which helps predict outcomes of unprecedented scenarios. The knowledge gained from data mining can also be subsequently used for different applications ranging from business management to medical diagnosis. Decision makers can hence make a more accurate assessment of situations based on this attained knowledge. Support vector machines (SVMs) have recently gained recognition as a powerful data mining technique to tackle the problem of knowledge extraction (Burges Christopher, 1998). SVMs use kernel functions to transform input features from lower to higher dimensions. Many practical applications exploit the efficiency and accuracy of SVMs, such as intrusion detection (Mukkamala, Janoski, & Sung, 2002) and bioinformatics where the input features are of very high dimensions.

Data mining is an essential step in the process of knowledge discovery in databases (KDD) (Fayyad, 1997). In addition to data mining, major steps of KDD also include data cleaning, integration, selection, transformation, pattern evaluation, and knowledge presentation. Since data is frequently interspersed with missing values and noise, which makes them incoherent, data preprocessing has thus become an important step before data mining to improve the quality of the data. This subsequently improves the data mining results. Data pre-processing takes several forms, including data cleaning, data transformation, and data reduction. Data cleaning is done to remove noise in the data. Data transformation is to normalize the data. Finally, data reduction is to reduce the amount of data by aggregating values or removing and clustering redundant attributes.

Removal of redundant attributes through selection of relevant attributes has become the focus of several recent search projects (Liu & Motoda, 1998). Several machine learning techniques have been around for attribute selection, including evolutionary algorithms (EAs), neural networks, and Bayes Theorem (Chang, Zheng, Wang, & Good, 1999; Hruschka & Ebecken, 2003; Mangasarian 2001; Tan et al., 2002; Wong, Lam, Leung, Ngan, & Cheng, 2000). Hruschka and Ebecken (2003) used the Bayesian approach to carry out attribute selection. The Markov Blanket of the class variable was used as a selection criterion. Neural networks and fuzzy logics (Benitez, Castro, Mantas, & Rojas, 2001) have also been employed

* Corresponding author.
E-mail address: eletankc@nus.edu.sg (K.C. Tan).

0957-4174/$ - see front matter © 2008 Published by Elsevier Ltd.
doi:10.1016/j.eswa.2008.10.013

for carrying out the attribute selection task. The attributes were first ranked according to a relevance measure. Attributes were then removed in an increasing order of relevance until the generalization ability of the network reached unacceptable levels. The downside of using neural networks is that they are not comprehensible to users. Furthermore, deciding the optimal number of neurons is a difficult task.

EAs appear to be promising in the field of attribute selection due to their heuristical nature in a directed, stochastic search. They are based on the process of natural selection and Darwin's theory of "*survival of the fittest*", which tend to drive an objective to an optimum. Recently, EAs have been applied in attribute selection for several applications (Martin-Bautista & Vila, 1999; Shi, Shu, & Liu, 1998). Pappa, Freitas, and Kaestner (2002) combined genetic algorithm (GA) and C4.5 (Quinlan, 1992) in a multiobjective approach. Multiobjective Genetic Algorithm (MOGA) was used to select the best attribute set by minimizing the error rate and the C4.5 tree size. The results derived demonstrated that the majority of the MOGA-found solutions dominated the baseline (the set of all attributes) and were distributed evenly along the Pareto front. This justifies the ability of GA to produce good results with a wide spread due to its randomness.

It is thus beneficial to investigate whether EAs and SVMs can be combined effectively to develop into a good classifier empowered by attribute selection. Based on the past successes of EAs and SVMs, they are fused in a hybrid approach to carry out both attribute selection and data classification. The workflow of this hybrid model contains two main stages. The first phase entails the selection of a set of attributes via EAs. These attributes are then passed to the SVM classifier to acquire a fitness measure for each attribute set in the second phase. These fitness values are then used in the selection of the best set of attributes based on GA. This cyclic method is known as the wrapper approach. Moreover, improvements are made by replacing unfit members of an existing population in a bid to increase the average fitness of the population and garner better results.

The remainder of the paper is organized as follows. Section 2 describes the attribute selection task in data mining and the approach used. Section 3 analyzes the proposed GA-SVM hybrid algorithm in the form of a flow chart. In addition, the main characteristics of the hybrid such as the chromosome structure, population layout, and the improved correlation-based algorithm are discussed. Section 4 presents the case study, which includes the introduction of experiment datasets and simulation results. The results are then tabulated and compared with several established algorithms. The viability and usefulness of the hybrid can be observed from the results and show its prospects for future data classification. Section 5 introduces the improvement of the proposed algorithm. Finally, section 6 presents the concluding remarks.

2. Attribute selection in data mining

2.1. Attribute selection

In the KDD process, interesting patterns and useful relationships are attained from the analysis of the input data. To ensure that the patterns derived are as accurate as possible, it is essential to improve the quality of the datasets in a pre-processing stage. Most real life data sets contain a certain amount of redundant data, which does not contribute significantly to the formation of important relationships. This redundancy not only increases the dimensionality of the data set and slows down the data mining process but also affects the subsequent classification performance. With this in mind, data reduction aims to trim down the quantity of data that is to be analyzed and yet produce almost similar, if not better, results as compared to the original data. More meaningful relationships can also be derived as the superfluous portions are removed.

Attribute selection is the process of removing the redundant attributes that are deemed irrelevant to the data mining task (Liu & Motoda, 1998). Seemingly, a ML algorithm's generalization ability improves with the number of attributes available. However, the presence of attributes that are not useful to classification might interfere with the relevant attributes to degrade classification performance. This is due to the noise that is contributed by these additional attributes and raises the level of difficulty of the ML algorithm in differentiating the signal from noise (Caruana & Freitag, 1994). Subsequently, the complexity of searching the attributes that produces good generalization is increased. The objective of attribute selection is therefore to search for a worthy set of attributes that produce comparable classification results to the case when all the attributes are used. In addition, a smaller set of attributes also creates less complicated patterns, which are easily comprehensible, and even visualizable, by humans.

The following step would be to find an algorithm that is efficient to carry out the search for the optimum and minimum set of attributes. It has to be noted that for a data set with n attributes, there are 2^n-1 possible subsets. Therefore, an exhaustive search for an optimal set of attributes would be time-consuming and computationally expensive if n is large. Several hill climbing methods have been investigated before, for example, the stepwise forward selection and stepwise backward elimination techniques. In forward selection, the search begins with an empty set and adds attributes with increasing relevance, before terminating at the point when the classification performance declines. Backward elimination starts with the complete set of attributes and prunes the most irrelevant attribute after each iteration. Due to the fact that forward selection begins with an empty set, it neglects the interaction between attributes, which may influence the selection process. On the other hand, backward elimination takes into account this interaction because it begins with a complete set of attributes. However, the analysis from the full set results in a lengthy runtime and may be unfeasible to carry out if the number of attributes is large. Another commonly used search method is best-first search (Ginsberg, 1993; Russell & Norvig, 1995), which is more robust than hill climbing. The major difference is that it is more exhaustive and evaluates the successors of the best attribute set in the solution space, unlike hill climbing which carries out exploration in a fixed path. This rigidity tends to lead the algorithm to a local optimum and terminates the search without achieving global optimality. Kohavi and John (1996) compared the hill climbing search with the best-first search for attribute selection and reported better results with the latter search. Despite this, the best-first search was similarly trapped in local optima in several of the artificial data sets tested.

In light of these findings, a more randomized approach would be more suitable to avoid the possibility of being confined in local optima. Hence, in this paper, a GA is used as the underlying search operator for attribute selection. Even if the algorithm arrives at a local optimum, the genetic operators would create opportunities to amend the situation. The stochastic nature of GA is the distinction that distinguishes it from the other searches. Empirical results also demonstrate a fast rate of convergence, which makes GA an efficient algorithm as the number of attributes n increases.

2.2. Wrapper vs filter approach

In the attribute selection process, there are two main approaches – the *wrapper*, and *filter* approach. The wrapper approach uses the actual data mining algorithm in its search for the attribute subsets (Kohavi & John, 1996) while in the filter approach, undesirable attributes are filtered out of the data before classification begins. Figs. 1 and 2 illustrate both methods.