Computers and Mathematics with Applications 58 (2009) 65–73

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

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

A hybrid approach to design efficient learning classifiers

Bikash Kanti Sarkar [a,*], Shib Sankar Sana [b]

[a] Department of Computer Science and Engineering, B.I.T., Mesra, Ranchi - 835 215, Jharkhand, India

[b] Department of Mathematics, Bhangar Mahavidyalaya (C.U.), Bhangar, Pin-743 502, 24-Pgs(S), W.B., India

ARTICLE INFO

Article history:
Received 3 February 2008
Received in revised form 17 December 2008
Accepted 12 January 2009

Keywords:
Rule discovery
Decision tree
Genetic algorithm
Efficient
Classification
Accuracy

ABSTRACT

Recently, use of a Learning Classifier System (LCS) has become promising method for performing classification tasks and data mining. For the task of classification, the quality of the rule set is usually evaluated as a whole rather than evaluating the quality of a single rule. The present investigation proposes a hybrid of the C4.5 rule induction algorithm and a GA (Genetic Algorithm) approach to extract an accuracy based rule set. At the initial stage, C4.5 is applied to a classification problem to generate a rule set. Then, the GA is used to refine the rules learned. Using eight well-known data sets, it has been shown that the present work, in comparison to C4.5 alone and UCS, provides a marked improvement in a number of cases.

1. Introduction

The idea of learning classification was first introduced in Holland in 1970. Classification concepts are important in the design of computerized information processing systems for several applications such as remote sensing, medical diagnosis, radar, etc. In order to implement a *multi-category* classification system, an efficient rule set is needed. Research in the rule induction field has been carried out for more than 30 years and has certainly produced a large number of algorithms. However, these are usually obtained from the combination of a basic rule induction algorithm with a new evaluation function. One of the biggest constraints in using some of the traditional machine learning methods for data mining is the problem of scaling up the methods to handle the huge size of the data sets and their high dimensionality. A survey of scaling up machine learning algorithms has been provided [1]. The use of Genetic Algorithms (GAs) [2] in addressing multi-category classification problems has been attempted by researchers in different ways.

Over the years, GAs have been successfully applied in learning tasks in different domains like chemical process control [3], financial classification [4], manufacturing scheduling [5], robot control [6] etc. A population of a fuzzy rule set [7] is evolved using a GP (Genetic Program: an extension of a GA) [8]. An accuracy based GA approach, UCS [9], is developed for performing the classification task. C4.5 (revision 8) [10] is one of the most successful and popular rule induction algorithms. In order to forecast the future sales of a printed circuit board factory more accurately, Chang et al. [11] have proposed a hybrid model in which a GA is utilized to optimize the Fuzzy Rule Base (FRB) adopted by the Self-Organization Map (SOM) neural network. In [12], the GA part of the hybrid model was employed to find an optimal structuring element for classifying garment defect types. Faraoun and Boukelif [13] made an attempt to show the use of a new GP classification approach for performing network intrusion detection. Wong et al. [14] proposed a decision support tool, combining an expert system and the Takagi–Sugeno Fuzzy Neural Network (TSFNN) for fashion coordination. They have also shown that the GA plays an important role here in reducing the number of coordination rules and the training time for TSFNN.

* Corresponding author.
E-mail addresses: bk_sarkarbit@hotmail.com (B.K. Sarkar), shib_sankar@yahoo.com (S.S. Sana).

0898-1221/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2009.01.038

In the present study, C4.5 as rule induction s/w (downloaded) extracts an initial rule set for any classification problem. The *interface* [15] takes the role of eliminating the 'IF–THEN' part from the generated rules, since rules in 'IF–THEN' form are not suitable for use in applying a genetic approach. The data sets used for this experiment are obtained from the UCI repository [16]. The continuous values of the attributes in the data sets are discretized using SPID4.7 [17] (a discretization algorithm). In the *post-processing* phase of the work, a GA method is included to optimize the set of rules learned by C4.5 in order to get high predictive accuracy. For the GA, a "fitness" function is required to evaluate the chromosome (offspring). The proposed algorithm combined with the fitness functions presented tries to retain the best rules with minimum classification *error* rate.

In this article, the learning capabilities of C4.5 and GA are combined to improve the performance of the classification problems and marked performance enhancement is shown.

The paper is organized as follows. Section 2 of this paper briefly discusses an overview of rule extraction systems. Section 3 gives some theoretical background for C4.5, genetic approaches to classification tasks and UCS. In Section 4, a brief description of the accepted methodology and findings along with comparison with an accuracy based approach (UCS) are presented. Section 5 gives our conclusion.

2. Rule extraction systems

Data mining [18–21] is very beneficial in economic and scientific domains. Techniques of Knowledge Discovery from a Database (KDD) are applied to reveal critical information hidden in data sets. Recently, the rule extraction task in KDD has attracted much attention to researchers. The goal of a rule extraction system is to induce a hypothesis (a rule set) from a set of *training* data samples.

A *classification* rule is a collection of some *non-target* attributes with a *target* (class) attribute. Each rule consists of an antecedent (condition) and a consequent (prediction/class): *antecedent* $\rightarrow$ *consequent*. The extracted model (a rule set, i.e., knowledge) can be applied in the prediction of *new instances*, unseen by the system. A rule extraction system includes the following components.

1. Data collection

Data are collected from various domains such as aerospace, banking and finance, retail, etc. For any classification problem, data may also be obtained from heterogeneous data sources.

Further, a specific data mining technique may be more appropriate to a particular data set. Grzymala-Busse et al. [22] have provided a comparison study of two data mining approaches from imbalanced data, and shown that the appropriate approach should be selected for specific imbalanced data sets.

2. Data preprocessing

The data to be operated on by the system may possess a number of variables (attributes). However, all the attributes are not necessary for analyzing data, i.e., some irrelevant attributes may be discarded from data sets. In order to remove irrelevant attributes, the following preprocessing of data is indispensable.

Feature selection: much research work [23,24] has been carried out on choosing a feature subset to represent the concept of data with the removal of irrelevant and redundant features.

Normalization: continuous and text inputs have to be converted into numerical form.

3. Selection of rule extraction tools

Neural networks, fuzzy logic, decision trees, rough sets, genetic algorithms etc. are used as *tools* for rule extraction.

Neural networks: are often referred to as artificial neural networks to distinguish them from biological neural networks. The neural network can be viewed as a directed graph with source (*input*), sink (*output*) and internal (*hidden*) nodes. The input nodes exist in an *input layer*, and the output nodes exist in an *output layer*. The hidden nodes occur in one or more hidden layers.

It is observed that in learning from previous experience, a neural network acts as an excellent tool in prediction. Many researchers focus on applying neural networks in the area of rule extraction [25–27]. However, a disadvantage is that it is difficult to design neural network architectures (mainly *activation functions* and *trained parameters*), and the results obtained may sometimes be incomprehensible. The study [28] demonstrates a new *neuron-adaptive* activation function and the use of AIS (Artificial Immune Systems) algorithms in extracting rules from trained neural networks.

Fuzzy logic: fuzzy sets have been applied in areas of computer science and databases specifically. In the classification problems, all records are assigned to one of the predefined classification regions. The *membership* function plays an important role in classification. Zadeh [29] introduces the concept of fuzzy logic and fuzzy set theory. Since then, several well-defined algorithms have been proposed by researchers. Wong et al. [30] discussed a fashion *mix-and-match* expert system (based on a fuzzy screening approach) which is developed to provide customers the *professional* and *systematic* mix-and-match recommendations automatically.

Decision trees: a decision tree [10,31] is a predictive modeling technique used in classification, clustering, and prediction tasks. It uses the *divide and conquer* technique to split the problem search space into subsets and extracts one rule (in IF–THEN form) for each leaf node of the tree. Also, it can form concise rules, in contrast to neural networks. In several domains, a decision tree based technique has been employed in performing classification tasks. For certain kinds of problems, better results may be achieved through this approach as compared to others.

Srdoc et al. [32] have shown that the decision tree based approach is suitable for estimating the possible *repair time* for a ship. Their research indicates also that standardization of problem domain notions and expertly designed databases