Chapter 4
A Fast Boosting Based Incremental Genetic Algorithm for Mining Classification Rules in Large Datasets

Periasamy Vivekanandan
Park College of Engineering and Technology, India

Raju Nedunchezhian
Kalaignar Karunanidhi Institute of Technology, India

ABSTRACT

Genetic algorithm is a search technique purely based on natural evolution process. It is widely used by the data mining community for classification rule discovery in complex domains. During the learning process it makes several passes over the data set for determining the accuracy of the potential rules. Due to this characteristic it becomes an extremely I/O intensive slow process. It is particularly difficult to apply GA when the training data set becomes too large and not fully available. An incremental Genetic algorithm based on boosting phenomenon is proposed in this paper which constructs a weak ensemble of classifiers in a fast incremental manner and thus tries to reduce the learning cost considerably.

INTRODUCTION

Genetic Algorithm (GA) based classification procedure (Spears & Gordon, 1993; Janikow, 1993; Greene & Smith, 1993; Yu, Goldberg, & Sastry, 2003; Rivera, 2004) is a stochastic search method which uses natural selection and reproduction techniques for examining and fine tuning a random population of candidate rules into a suitable, accurate solution for the given problem. The accuracy of the rules that GA finds are comparable and some times even more accurate than the rules obtained by other classification algorithms (Yang, Dwi, Widyantoro, Ioerger, & Yen, 2001) and it also performs well in complex domains where other methods fails . This is due to the ability of GA to build the model based only on natural evolution process and not based on the Domain knowledge. But when the size of the example data set increases, GA becomes unacceptable because of its high learning cost.

DOI: 10.4018/978-1-4666-3628-6.ch004

This is due to the repeated access of the data set by the Algorithm for evaluating its candidate rule set accuracy. This behavior should be minimized in order to make it scalable to large data sets.

PREVIOUS WORK

Boosting is one of the commonly used classifier learning approach (Treptow & Zell, 2004; Freund & Schapire, 1995; Freund & Schapire, 1996; Schapire & Singer, 1998). According to Schapire and Singer (1998) boosting is a method of finding a highly accurate hypothesis by combining many "weak" hypotheses, each of which is only moderately accurate. It manipulates the training examples to generate multiple hypotheses. In each iteration, the learning algorithm uses different weights on the training examples, and it returns a hypothesis ht. The weighted error of ht is computed and applied to update the weights on the training examples. The result of the change in weights is to place more weight on training examples that were misclassified by ht, and less weight on examples that were correctly classified. The final classifier is constructed by weighted vote of the individual classifiers. In the proposed method many weak GA based classifiers are built iteratively. When combined, the weak classifiers form an accurate strong model because ensemble classifiers outperform single classifiers (Chandra & Yao, 2006).

In the GA literature Complexity arising due to Scalability is mainly addressed by parallel processing.(*Wilson Rivera, 2004*; Lopes & Freitas, 1999) For example parallel genetic algorithm proposed by Lopes and Freitas (1999), addresses the scalability issue with respect to GA. It involves multiple processors and the data set is divided into multiple parts (data sets). The multiple data sets are distributed to multiple processors and each processor generates rules for each data set. The rules generated by each processor are again shared by all processors for fitness calculation.

Incremental learning is very popular in making the classification methodologies (Domingos & Hulten, 2000; Spencer & Domingos, 2001; Polikar, Upda, Upda, & Honavar, 2001; Gao, Ding, Fan, Han, & Yu, 2008) scalable and they try to build the model by scanning the data set only once. Only very few methods in the data mining literature employs GA as base algorithm for incremental learning. An incremental GA was proposed by Guan and ZhuCollard (2005), for a dynamic environment which updates the rules based on the new data. Due to the arrival of new data or new attribute or class, the classification model may change. So to deal with this the author proposes an incremental based GA.

PROPOSED METHOD

Overview

In genetic algorithm the potential solution is represented by individuals called candidate rules (Dehuri & Mall, 2006) and are of the form

$A1_{\wedge}$ A2…An THEN C.

The antecedent part of the rule is the conjunction of conditions say A (conjunction of attribute value pairs A1, A2…An) and the consequent part C is the class label.

These candidates are initially generated randomly and are processed in Sequential steps, such that an accurate solution gradually emerges. The Sequential Steps are called generations. Each generation goes through the selection (or testing) phase and the reproduction phase. During the selection phase, candidate rules accuracy (fitness) is evaluated using the example data set. The Fitness of the rules is calculated based on a function containing two terms namely predictive accuracy and comprehensibility (Dehuri & Mall, 2006). A very simple way to measure the predictive accuracy is