Chapter 13
Evolution of Genetic Algorithms in Classification Rule Mining

Dipankar Dutta
University Institute of Technology, The University of Burdwan, India

Jaya Sil
Bengal Engineering and Science University, India

ABSTRACT

Classification is one of the most studied areas of data mining, which gives classification rules during training or learning. Classification rule mining, an important data-mining task, extracts significant rules for classification of objects. In this chapter class specific rulesare represented in IF <Antecedent> THEN <Consequent> form. With the popularity of soft computing methods, researchers explore different soft computing tools for rule discovery. Genetic algorithm (GA) is one of such tools. Over time, new techniques of GA for forming classification rules are invented. In this chapter, the authors focus on an understanding of the evolution of GA in classification rule mining to get an optimal rule set that builds an efficient classifier.

INTRODUCTION

Data mining (DM) is to extract a priori unknown, valid and actionable information from different types of data sets. Using the information crucial decisions are made. Major DM tasks are classification, clustering, association rule discovery, sequential pattern discovery, regression, deviation detection, prediction, data visualization etc. One of the most important tasks in DM is mining classification rules from data sets. In a classification problem, records are assigned to one class among a small set of prespecified classes. Application domains include medical diagnosis, fault detection in electromechanical devices, evaluating credit applications, predicting crop quality etc. (Corcoran and Sen, 1994). Some good literatures on DM are (Piatetsky-Shapiro and Frawley, 1991; Fayyad et al., 1996; Weiss and Indurkhya, 1998; Westphal and Blaxton, 1998; Pyle, 1999; Freitas, 2002; Mitra et al., 2002; Dunham, 2003; Mitra and Acharya, 2003; Witten and Frank, 2005; Tan et al., 2005; Han and Kamber, 2006).

DOI: 10.4018/978-1-4666-2518-1.ch013

Learning classifier employ two biological metaphors; evolution and learning... [where] learning guides the evolutionary unit to move towards getting a better set of rules (Dam et al., 2008). As evolutionary approach GAs are applied to design an efficient rule base classifier and a learning mechanism suitable for the problem. However, GAs and learning mechanism are dependent on the training data.

GAs are "search algorithms based on the dynamics of natural selection and natural Genetics" (Goldberg, 1989). Broad categories of GAs are - single objective GAs and multi objective GAs (MOGA). When an optimization problem involves only one objective, the task of finding the best solution is a single objective optimization problem. Integer programming, dynamic programming, geometric programming, stochastic programming and various other methods are available to solve these types of problems. But real world problems are more complex. Many objectives are to be minimized or maximized simultaneously in most of them. These are multi objective optimization problem (Deb, 2001). The use of GAs in classification rule mining is an attempt to exploit effectively the large search space (Fidelis et al., 2000) by proving better interactions among the features. Almost all conventional methods search from a single point and converge to a solution. It is sensitive to the choice of first solution. GAs always work on a whole population of points (strings) and perform a nearly global search rather than performing a local, hill-climbing search. This contributes much for robustness of GAs. It improves the chance of reaching the global optimum or reduces the risk of being trapped in a local optimum point. The knowledge discovery models mostly used in DM is yet based on rule induction methods, which resort to local search (Fidelis et al., 2000).

In course of machine learning, machine learns from training data through some learning mechanism. In artificial intelligence, learning is "the improvement of performance in some environment through the acquisition of knowledge resulting from experience in that environment" (Langley, 1995). Broad categories of learning are i) Supervised learning ii) Reinforced learning and iii) Unsupervised learning. In DM problems all training data are presented simultaneously to the learner resulting in offline or batch learning. Alternatively, online or incremental learning implies that training instances are presented to the learner one at a time (Langley, 1995; Ben-David et al., 1997; Harries et al., 1998).

The main objective of this chapter is to discuss the technical evolution of the GAs in classification rule mining. We have provided a brief survey of different classification techniques and a detailed survey on classification rule mining using GAs in the chapter.

LITERATURE REVIEW

The classification techniques and specifically rule discovery to build the classifiers are broadly categorized as soft computing based methods and non-soft computing based methods.

Non-Soft Computing Based Methods

We are grouping the methods as follows.

Tree Based Methods

These methods build one inverted tree by dividing data based on some heuristic measures, e.g. information gain and information gain ratio. This method is repeated from the root node of the tree considering all training data unless class labels are assigned to every leaf node. Each branch of the tree from root to leaves represents rules. State-of-the-art classification methods C4.5 (Quinlan, 1993) belong to this category. Others are GOTA (Hartmann et al., 1982) and CART (Breiman et al., 1984).