IGI PUBLISHING
ITJ3905
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-pub.com

This paper appears in the publication, **International Journal of Data Warehousing and Mining**, Volume 3, Issue 4
edited by **David Taniar© 2007, IGI Global**

Evolutionary Induction of Mixed Decision Trees[1]

Marek Kretowski, Bialystok Technical University, Poland

Marek Grzes, Bialystok Technical University, Poland

ABSTRACT

This article presents a new evolutionary algorithm (EA) for induction of mixed decision trees. In non-terminal nodes of a mixed tree, different types of tests can be placed, ranging from a typical inequality test up to an oblique test based on a splitting hyper-plane. In contrast to classical top-down methods, the proposed system searches for an optimal tree in a global manner, that is it learns a tree structure and finds tests in one run of the EA. Specialized genetic operators are developed, which allow the system to exchange parts of trees, generating new sub-trees, pruning existing ones as well as changing the node type and the tests. An informed mutation application scheme is introduced and the number of unprofitable modifications is reduced. The proposed approach is experimentally verified on both artificial and real-life data and the results are promising. Scaling of system performance with increasing training data size was also investigated.

Keywords: *decision trees; evolutionary algorithms; global induction; mixed decision trees*

INTRODUCTION

Decision trees (Murthy, 1998) are one of the most frequently applied data mining approaches. There exist many induction algorithms, which tackle the problem of building decision trees in a different way. Most frequently, they differ in the measure for the test assessment, but also in the type of search in solution space (i.e., top-down vs. global). From a user's point of view, one of the most important features of a decision tree is a test representation in the internal nodes. In typical univariate trees, two types of tests are usually permitted. For a nominal attribute, mutually exclusive sets of feature values are associated with each branch, whereas for a continuous valued feature inequality tests are applied. In the case of multivariate trees, more than one feature can be used to create a test. Oblique tests based on a splitting hyper-plane are the most widely used form of multivariate tests. Most of the DT-based systems are homogeneous, which means that they take advantage of only one type of test (i.e., univariate or oblique). *C4.5* (Quinlan, 1993) can be treated as one of the best-known representatives of the first type, whereas *OC1* (Murthy, Kasif, & Salzberg, 1994) is a good example of an oblique tree inducer.

These two systems belong to the group of decision tree algorithms, which represent the de-facto standard for empirical evaluations and are commonly used for comparisons.

The term *mixed decision trees* was proposed by Llora and Wilson (2004) to describe trees in which different types of tests can be exploited. One of the first and best-known examples of such an approach is the *CART* system (Breiman, Friedman, Olshen, & Stone, 1984). This system is able to search for a linear combination of non-nominal features in each node and it compares the obtained test with the best univariate test. However, it should be noted that *CART* has a strong preference for simpler tests; it rarely uses the more elaborate splits. Another form of a hybrid classifier is proposed by Brodley (1995). Her *MCS* system combines univariate tests, linear machines, and instance-based classifiers (*k-NN*) and during the top-down generation of a tree classifier it recursively applies automatic bias selection. Recently, a fine-grain parallel model *GALE* (Llora et al., 2004) was applied to generate decision trees, which employ inequality and oblique tests.

There are two main approaches to the decision tree induction: top-down and global. The first one is based on a greedy recursive procedure of test searching and sub-node creation until a stopping condition is met. The locally optimal tests according to the predefined criteria are chosen in each step, but such a procedure does not guarantee the global optimality of the final tree. This problem can be easily observed when there is a strong interaction between features. Only treating them together can lead to the optimal solution. Additionally, the post-pruning is usually applied after the actual top-down induction to avoid the problem of over-fitting the training data. It should be noted that post-pruning techniques have only limited ability to correct the tree structure. The *C4.5* and *OC1* systems apply the top-down approach and are used for the comparison in this article.

In contrast to the classical top-down approach, global algorithms try to simultaneously search for both the tree structure and all tests in non-terminal nodes. This process is obviously much more computationally complex but it can reveal hidden regularities, which are almost undetectable by greedy methods. The global induction is mainly represented by systems based on evolutionary approach.

Evolutionary computations (Michalewicz, 1996) are stochastic techniques, which have been inspired by the process of biological evolution. Their success is attributed to the ability to avoid local optima, which is their main advantage over greedy search methods. Evolutionary techniques are known to be useful in many data mining tasks (Freitas, 2002). They were successfully applied in the framework of both top-down and global systems to learning univariate (Fu, Golden, Lele, Raghavan, & Wasil, 2003; Koza, 1991; Nikolaev & Slavov, 1998; Papagelis & Kalles, 2001) and oblique trees (Bot & Langdon, 2000; Chai, Huang, Zhuang, Zhao, & Sklansky, 1996; Cantu-Paz & Kamath, 2003; Kretowski, 2004).

The global approach based on evolutionary algorithms for decision tree induction was investigated in our previous articles. We showed that homogeneous trees, univariate (Kretowski & Grzes, 2005a) or oblique (Kretowski & Grzes 2005b, 2006) can be effectively induced and we demonstrated that globally generated classifiers are generally less complex with at least comparable accuracy. In this article, we want to merge the two developed methods in one system, which will be able to induce mixed trees.

The rest of the article is organized as follows. In the next section our global system for induction of mixed decision trees is presented. Experimental validation of the approach on both artificial and real-life datasets is presented in the third section. The article finishes with our conclusion.

GLOBAL INDUCTION OF MIXED DECISION TREES

The algorithm proposed in this article applies a global approach to decision tree induction based on evolutionary computation. The general structure of the proposed solution follows a typical evolutionary framework (Michalewicz, 1996).