Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 183 (2007) 767–784
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Computing, Artificial Intelligence and Information Management

MEPAR-miner: Multi-expression programming for classification rule mining

Adil Baykasoğlu [a,*], Lale Özbakir [b]

[a] University of Gaziantep, Department of Industrial Engineering, 27310 Gaziantep, Turkey
[b] Erciyes University, Department of Industrial Engineering, Kayseri, Turkey

Received 13 April 2006; received in revised form 13 October 2006; accepted 13 October 2006
Available online 13 December 2006

Abstract

Classification and rule induction are two important tasks to extract knowledge from data. In rule induction, the representation of knowledge is defined as IF-THEN rules which are easily understandable and applicable by problem-domain experts. In this paper, a new chromosome representation and solution technique based on Multi-Expression Programming (MEP) which is named as MEPAR-miner (*Multi-Expression Programming for Association Rule Mining*) for rule induction is proposed. Multi-Expression Programming (MEP) is a relatively new technique in evolutionary programming that is first introduced in 2002 by Oltean and Dumitrescu. MEP uses linear chromosome structure. In MEP, multiple logical expressions which have different sizes are used to represent different logical rules. MEP expressions can be encoded and implemented in a flexible and efficient manner. MEP is generally applied to prediction problems; in this paper a new algorithm is presented which enables MEP to discover classification rules. The performance of the developed algorithm is tested on nine publicly available binary and *n-ary* classification data sets. Extensive experiments are performed to demonstrate that MEPAR-miner can discover effective classification rules that are as good as (or better than) the ones obtained by the traditional rule induction methods. It is also shown that effective gene encoding structure directly improves the predictive accuracy of logical IF-THEN rules.

Keywords: Data mining; Classification rules; Multi-expression programming; Evolutionary programming

1. Introduction

There is a tremendous increase in the volume of data in every business and scientific arena. In order to turn these massive data into useful information there has been a growing interest in the area of knowledge discovery (Baykasoğlu and Özbakir, 2005). Data Mining (DM) can be considered as one of the steps in the knowledge discovery process. It is applied to many different problem domains with considerable success. The algorithms and tools of DM are coming from variety of disciplines like artificial intelligence, statistics, data base systems etc as it's shown in Fig. 1. Most of the algorithms and techniques are still in the development stage.

DM algorithms are developed for different types of applications. Therefore, they aim to reach

* Corresponding author. Tel./fax: +90 342 3604383.
E-mail address: baykasoglu@gantep.edu.tr (A. Baykasoğlu).

0377-2217/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2006.10.015

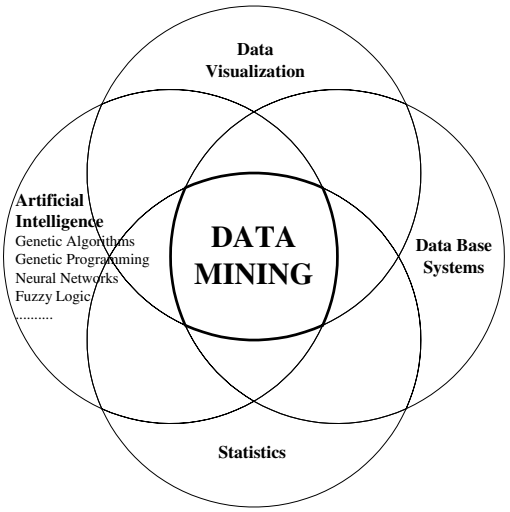

Fig. 1. Data mining and related fields.

different types of knowledge. DM algorithms are grouped into six categories by Chen et al. (1996). Mining of association rules, data generalization and summarization, classification, data clustering, pattern based similarity search, mining path traversal patterns.

Classification is one of the most commonly applied supervised data mining technique. Classification formulates a classification model based on data. The model can be used to classify a new record of data into one of the predefined classes based on the values of attributes (Wong and Leung, 1999). Depending on the information available on classes and the type of classification, solution approaches are distinguished such as decision trees, rule induction, neural networks, *K*-nearest neighbors, Bayesian methods, evolutionary algorithms etc. Evolutionary algorithms have certain advantages which make them suitable for application in classification. The main motivation for using evolutionary algorithms in the discovery of prediction rules is that they perform a global search and can efficiently cope with interactions between attributes. Furthermore, they are convenient for parallelization of the data mining process, which can improve the efficiency of computation (Takac, 2003).

Evolutionary algorithms are defined as randomized search procedures inspired by the working mechanism of genetics and natural selection. There are different types of evolutionary algorithms such as genetic algorithms (GA), genetic programming (GP), evolution strategies (ES), and evolutionary programming (EP). These algorithms are based on the same concepts but differ in the way they represent solutions and on the operators used to create next generation.

Genetic programming is the iterative application of Darwinian principles of adaptation by natural selection and survival of the fittest to populations of computer programs (Koza, 1992). GP is an extension of GA. The main difference between them is the representation of the structure they manipulate on and mechanics of the operators applied on this structure. Genetic programming based approaches, which are also classified under the artificial intelligence based techniques, are becoming popular in complex data mining applications. They are generally applied to prediction problems but they are also very suitable for classification problems, because GP is potentially capable of providing a solution structure between two extreme positions in data classification. One of these extreme points is neural networks which provide accurate classifiers but work as a black box. The other one is classification algorithms such as C4.5 which can produce excessively complex decision trees.

GP can be considered as a more open-ended search concept than GA. The search performed by GP can be very useful in classification tasks since GP can produce many different combinations of attributes. Freitas (2002) presented a survey of evolutionary algorithms, particularly genetic algorithms and genetic programming in data mining and knowledge discovery. He mainly focused on classification type problems. Freitas (1997) also proposed a GP framework for classification and generalized rule induction. Carvalho and Freitas (2002a,b) proposed a hybrid decision tree/genetic algorithm method for discovering classification rules. Decision trees produced by C4.5 algorithm are used to discover large disjuncts. In the second phase GA is used to discover rules covering the examples belonging to small disjuncts. Takac (2003) combined GP algorithm with the cellular parallel model of genetic algorithms. The proposed approach is applied to the credit classification and heart disease classification tasks. Zhou et al. (2003) presented a new approach for evolving classification rules through Gene Expression Programming which was initially proposed by Ferreira (2001) as a linear genetic programming approach. They tested their approach by using 12 data sets collected from the literature. De Falco et al. (2002) developed a genetic programming framework which is capable of performing automatic discovery of