Available online at www.sciencedirect.com
ScienceDirect
Applied Mathematics and Computation 195 (2008) 786–798
APPLIED MATHEMATICS AND COMPUTATION
www.elsevier.com/locate/amc

Mining classification rules with Reduced MEPAR-miner Algorithm

Emel Kizilkaya Aydogan [a], Cevriye Gencer [b,*]

[a] Department of Industrial Engineering, Faculty of Engineering, Erciyes University, Kayseri, Turkey
[b] Department of Industrial Engineering, Faculty of Engineering and Architecture, Gazi University, Ankara, Turkey

Abstract

In this study, a new classification technique based on rough set theory and MEPAR-miner algorithm for association rule mining is introduced. Proposed method is called as 'Reduced MEPAR-miner Algorithm'. In the method being improved rough sets are used in the preprocessing stage in order to reduce the dimensionality of the feature space and improved MEPAR-miner algorithms are then used to extract the classification rules. Besides, a new and an effective default class structure is also defined in this proposed method. Integrating rough set theory and improved MEPAR-miner algorithm, an effective rule mining structure is acquired. The effectiveness of our approach is tested on eight publicly available binary and *n*-ary classification data sets. Comprehensive experiments are performed to demonstrate that Reduced MEPAR-miner Algorithm can discover effective classification rules which are as good as (or better) the other classification algorithms. These promising results show that the rough set approach is a useful tool for preprocessing of data for improved MEPAR-miner algorithm.

Keywords: Data mining; Classification rules; Attribute reduction; Rough set; Evolutionary programming

1. Introduction

Knowledge Discovery in Databases (KDD) has become a very attractive discipline both for research and industry within last few years. Its goal is to extract pieces of knowledge or 'patterns' from usually very large databases [1].

Rough set methodology provides a powerful tool for knowledge discovery from large and incomplete sets of data. A number of algorithms and systems have been developed based on the rough set theory [2]. Babu et al. [3] proposed a new learning approach integrating the activities of data abstraction, frequent item generation, compression, classification and the use of rough sets. Hassan and Tazaki [4] combined rough set theory and Genetic programming (GP) for deriving knowledge rules from medical database. Similarly, hybrid algorithms based on neural network and rough set theory are introduced in literature [5–7]. The main idea

* Corresponding author.
E-mail address: ctemel@gazi.edu.tr (C. Gencer).

0096-3003/$ - see front matter
doi:10.1016/j.amc.2007.05.024

in these studies is accelerating and simplifying the process of using neural networks for mining knowledge. In addition, Pal and Mitra [8] introduced a new hybrid algorithm based on fuzzy sets and rough set theory for case generation, and showed efficiency of this algorithm on real-life data sets. Similarly, Wong et al. [9] proposed a method based on the concepts of Genetic Algorithm (GA) and SVD-QR method to construct an appropriate fuzzy system for pattern classification. Huang et al. [10] proposed a hybrid approach of rough set theory and genetic algorithm for fault diagnosis. Zhang et al. [11] introduced a hybrid classifier based on rough set theory and support vector machines called RS-SVMs to recognize radar emitter signals.

On the other hand, MEPAR-miner [12] is a new algorithm known as 'Multi-Expression Programming for Association Rule Mining' which is based on genetic programming and one of the most successful algorithm in association rule mining literature. But there are some drawbacks of this algorithm. Firstly, in the encoding structure, all of the attributes are used. Thus, the search space dimensions also get larger and the possibility of getting trap of local optima increases. Secondly, in the MEPAR-miner algorithm, a simple default class structure which depends on the mostly encountered class is used. It is observed that, on the test data no meaningful effect exists in the predictive accuracy operations of such default class structure. Finally, in the MEPAR-miner algorithm, since chromosome objective value calculation is time consuming (because of the reason that; for every terminal and function, all the training data sets are considered), solution time of simple genetic algorithm that is used may take a very long time. In order to eliminate these disadvantages, a Reduced MEPAR-miner Algorithm which has an effective encoding and default class structure and works according to parallel steady state genetic algorithm logic was improved.

This paper is organized as follows: In the first two sections, a brief description about rough set theory and MEPAR-miner algorithm is explained. The next section introduces our new detailed method. Following section presents some illustrative applications of our approach. Last section concludes our paper.

2. Rough set theory

Rough Set Theory (RST), introduced by Pawlak in the early 1980s, is a mathematical tool to deal with classification problems. It is based on the assumption that data and information are associated with every object of the universe of discourse. According to the definition given in Pawlak [13], a knowledge representation system or an information system is a pair $S = (U, A)$, where U is a non-empty, finite set of objects (called the universe), and A is a non-empty set of attributes.

The RS theory is based on the observation that objects may be indiscernible because of limited available information. For a subset of attributes $B \subseteq A$, the indiscernibility relation is defined by $IND(B)$ [13]:

$$IND(B) = \{(x, y) \in U^2 | a \in Ba(x) = a(y)\}.$$

$IND(B)$ is an equivalence relation on the set U. The relation $IND(B)$, $B \subseteq A$, constitutes a partition of U which is denoted $U/ND(B)$. If $(x, y) \in IND(B)$, then x and y are indiscernible by attributes from B. The equivalence classes of the B-indiscernibility relation are denoted $[X]_B$. For a subset $X \subseteq U$, the P-lower approximation of X can be defined as

$$\underline{B}X = \{x | [x]_{\mathrm{B}} \subseteq X\}.$$

Let $IND(B)$ and $IND(Q)$ be indiscernibility relations on U defined by the subset of attributes $B \subseteq A$ and $Q \subseteq \mathrm{A}$ respectively. An often-applied measure is the dependency degree of Q on B, which is defined as follows [13]:

$$\gamma_B(Q) = \frac{|POS_B(Q)|}{|U|}.$$

$POS_{\mathrm{B}}(\mathrm{Q}) = \cup_{X \in U/IND(Q)} \underline{B}X$, called a positive region of the partition $U/IND(Q)$ with respect to B.

One of the major applications of rough set theory is the attribute reduction that is the elimination of attributes considered to be redundant while avoiding information loss. The reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. Using the dependency degree as a measure, attributes are removed so that the reduced set provides the same dependency degree as the original. In a decision system, a reduct is formally defined as a subset R of the conditional attribute set X such that