Applied Soft Computing 37 (2015) 456–466

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Rule induction for hierarchical attributes using a rough set for the selection of a green fleet

Shian-Hua Lin[a,1], Chun-Che Huang[b,*], Zhi-Xing Che[b,2]

[a] Department of Computer Science and Information Engineering, National Chi Nan University, Taiwan #1, University Road, Pu-Li 545, Taiwan
[b] Department of Information Management, National Chi Nan University, Taiwan #1, University Road, Pu-Li 545, Taiwan

ARTICLE INFO

Article history:
Received 16 June 2013
Received in revised form 6 August 2015
Accepted 7 August 2015
Available online 22 August 2015

Keywords:
Green fleet selection
Rough set theory
Concept hierarchy
Decision rules

ABSTRACT

Rough set theory (RST) has been the subject of much study and numerous applications in many areas. However, most previous studies on rough sets have focused on finding rules where the decision attribute has a flat, rather than hierarchical structure. In practical applications, attributes are often organized hierarchically to represent general/specific meanings. This paper (1) determines the optimal decision attribute in a hierarchical level-search procedure, level by level, (2) merges the two stages, generating reducts and inducting decision rules, into a one-shot solution that reduces the need for memory space and the computational complexity and (3) uses a revised strength index to identify meaningful reducts and to improve their accuracy. The selection of a green fleet is used to validate the superiority of the proposed approach and its potential benefits to a decision-making process for transportation industry.

1. Introduction

Rough set theory (RST) was developed by Pawlak [48] to classify imprecise, uncertain, or incomplete information or knowledge that is expressed using data that is acquired from experience. The main advantage of rough set theory is that it does not require any preliminary or additional information about data: such as statistical probability, the basic probability assignment that is required for the Dempster–Shafer theory, or the grade of membership or the value of the possibility that is required for fuzzy set theory [28], so it has become the subject of an increasing number of studies and applications in many areas.

However, most of the previous studies on rough sets have focused on finding rules where a decision attribute is not hierarchical, but on the same level. In real-world applications, attributes are usually predefined hierarchically and can be represented by a conceptual hierarchy using hierarchical trees [35]. Some RS methods use two stages to generate reducts and then induce decision rules, e.g., a reduct extraction algorithm (REA) and an alternative reduct extraction algorithm (AREA) [44]. A large computational space is required to store the reducts at the first stage and the search for a solution is complex. In previous studies, the reducts often are often compared using the strength index (SI), which was introduced to identify meaningful reducts [13]. However, the use of a SI is limited to the same number of condition attributes and the results may not provide accurate hierarchical attributes. SI is a method of averaging the weights of attributes that are selected in the reducts.

To address the drawbacks of previous studies, the method proposed by this paper achieves the following objectives:

(1) The two stages (generating reducts and inducting decision rules) are merged, to reduce memory space and computational complexity.
(2) A revised strength index is used to identify meaningful reducts from all of the reducts, rather than from a part of the attributes that are selected in the reducts.
(3) The most specific decision attribute is determined, level by level, in the level-search procedure.

The remainder of this study is organized as follows: Section 2 surveys the literature related to rough set theory. The HRS problem and the level-search procedure are detailed in Section 3. The proposed method is detailed in Section 4. A case from the transportation industry is addressed in Section 5, to show how the decision rules are inducted and to validate the superiority of the proposed approach. The decision rules that are inducted by the proposed method allow netter decision-making processes for the purchase of new fleets. Section 6 provides concluding remarks.

* Corresponding author. Tel.: +886 492910960x4655; fax: +886 492915205.
E-mail addresses: shlin@ncnu.edu.tw (S.-H. Lin), cchuang@ncnu.edu.tw (C.-C. Huang), livebreeze@gmail.com (Z.-X. Che).
[1] Tel.: +886 492910960x4869.
[2] Tel.: +886 492910960x4442.

http://dx.doi.org/10.1016/j.asoc.2015.08.016
1568-4946/

2. Literature review

Rough set theory (RST) was developed by Pawlak [48] to classify imprecise, uncertain, or incomplete information or knowledge that is expressed using data that is acquired from experience [48]. In RST, a reduct is the minimal subset of attributes that allows the same classification of elements of the universe as the whole set of attributes. In RST, attributes are classified into two sets: condition and decision attributes. The latter refers to outcomes of the data set, in the form of condition attributes. Rough set theory has proved to be an excellent mathematical tool for the analysis of a vague description of objects (called actions in decision problems). The adjective, "vague", refers to the quality of information and defines the inconsistency or ambiguity that follows from information granulation [47,48]. RST has become the subject of much study and many applications in many areas, such as parallel computing [26], supply chain management [11,22,33,46], wave height prediction [1], text categorization [32], the prediction of debris flow disasters [30], rural development [3], webpage classification and Web services composition [9,25], electronic noses [2], IC packaging foundry [17], linguistic terms [16], elementary education [43], environmental performance evaluation [29], medical science [6,8,23,41], economic and financial prediction [9,42], the airline market [24], customer relationship management [37,39], transportation [4,20] and other real-life applications [12,18,19]. RST is successful because: (i) only the facts that are hidden in data are analyzed, (ii) no additional information about the data is required, such as thresholds or expert knowledge, and (iii) a minimal knowledge representation can be attained [31].

However, previous studies on RST have only focused on finding specific rules and possible rules when the decision attributes are on the same level, rather than hierarchical levels. In real-world applications, hierarchical attributes are usually predefined and can be represented by a conceptual hierarchy using hierarchical trees [35]. A concept hierarchy is a concise and general form of conceptual description that organizes the relationships between data [40]. Tseng et al. [36] proposed an approach to generate conceptual hierarchies, given a data set with nominal attributes that are based on a rough set. Yang et al. [40] built a simple higher-level decision table that used a conceptual hierarchical tree. Dong et al. [15] presented a model and a method for hierarchical fault diagnosis for a substation using rough set theory. The approach not only improved the efficiency of the discovery process, but also expressed the user's preference for guided generalization. However, these studies did not consider a decision attribute that has combinations of different attribute-level values, for example, outcome O1 at level 1, outcome O2 at level 2 and O3 at level 2.

Traditional RS method also cannot produce rules that allowed an ordered preference, so they cannot produce more meaningful and general rules [14]. Induction that uses RST often generates redundancy rules and cannot guarantee a credible classification of a decision table, as demonstrated in previous studies, such as the generation of classification rules in [7], the use of information-rich data to reduce data redundancy in [31], the analysis of diabetic databases in [21], the consistency and completeness of a nutrition management model in [27], the prediction of debris flow disasters in [30], or an alternative methodology to search for rules for large-scale data sets [46]. Tseng [38] proposed a new RST method, called an alternative rule extraction algorithm (AREA), which discovers preference-based rules by using the reducts with the maximum strength index (SI). This method identifies meaningful reducts, but it uses two stages to generate reducts and induct decision rules. A large computing space is required to store the reducts from the first stage. The procedure for searching for a solution is complex and the reducts are compared using the SI, which is limited to the particular condition attributes that are selected in some reducts.

Based on the literature review, the hierarchical rough set (HRS) problem in traditional RS application is defined and a new level-search procedure is proposed in Section 3.

3. The hierarchical rough set (HRS) problem and the level-search procedure

In this section, the structure of a conceptual hierarchy is constructed to represent the hierarchical attributes in Section 3.1. The hierarchical rough set (HRS) problem is defined in Section 3.2 to define the weaknesses of a traditional RS approach in practical applications. The level-search procedure is proposed in Section 3.3 to demonstrate how the final decision rules are produced.

3.1. The structure of the conceptual hierarchy

In this study, the concept hierarchy in the decision attributes is considered.

Notation:

O	a decision attribute set;
k	hierarchical index of a decision attribute;
l	level index;
e	entry point;
Hk	the concept hierarchy corresponding to O;
Hk_l	the concept hierarchy corresponding to O at level l;
sk	number of levels in the Hk;

If a conceptual hierarchy Hk refers to a set of domains, $Ox, \ldots, Oz$. $Hk_{sk}:\{Ox \times \ldots \times Oz\} \to Hk_{sk-1} \to Hk_1$, where Hk_{sk} denotes the set of concepts at the skth level, Hk_{sk-1} denotes the concepts at one level higher than those at Hk_{sk} and Hk_1 represents the top level, which is denoted as "ANY". O_x refers to the value of attribute = x at level 1, $O_{x.y}$ refers to the value of attribute = x at level 1 and that of y at level 2 and $O_{x.y.z}$ refers to the value of attribute = x at level 1, y at level 2 and z at level 3.

$O_{x.y.z} \in O_{x.y} \in O_x$ implies that a rule in the lower level has more specific information. In the conceptual tree, O_x is one of the parent nodes of $O_{x.y}$, $O_{x.y}$ is a child node of O_x and $O_{x.y}$ and $O_{x.n}$ are sibling nodes. In the hierarchical conceptual tree, each node represents a concept. The most general and universal concept is represented by the node at the top level. The specific concepts are represented by the node at a low level. In Fig. 1, for example, $O_{A.1} \cup O_{A.2} = HA_2 \in O_A$, $O_{B.1} \cup O_{B.2} = HB_2 \in O_B$, $O_A \cup O_B = H_1 \in O_0$.

3.2. Hierarchical rough set (HRS) problem

The hierarchical rough set problem is defined as:
Given:

- An information system, $I = (U, A)$, where U is a finite set of objects and A is a finite set of attributes, the elements of A are called condition attributes.

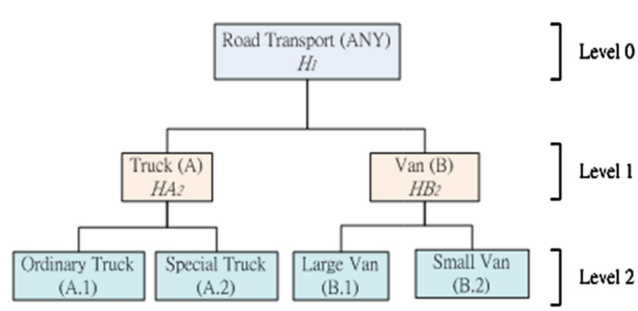

Fig. 1. The conceptual tree for a land transportation hierarchy.