



Alternative rule induction methods based on incremental object using rough set theory

Chun-Che Huang^a, Tzu-Liang (Bill) Tseng^b, Yu-Neng Fan^{c,*}, Chih-Hua Hsu^a

^a Department of Information Management, National Chi Nan University, No. 1, University Road, Puli, Nantou 545, Taiwan, ROC

^b Department of Industrial Manufacturing and Systems Engineering, The University of Texas at El Paso, 500 West University Avenue, El Paso, TX 79968, United States

^c Department of Information Management, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei City 106, Taiwan, ROC

ARTICLE INFO

Article history:

Received 1 January 2012

Received in revised form 1 July 2012

Accepted 25 August 2012

Available online 1 September 2012

Keywords:

Rough set theory

Incremental algorithm

Incremental object

Rule induction

Mining methods and algorithms

ABSTRACT

The rough set (RS) theory can be seen as a new mathematical approach to vagueness and is capable of discovering important facts hidden in that data. However, traditional rough set approach ignores that the desired reducts are not necessarily unique since several reducts could include the same value of the strength index. In addition, the current RS algorithms have the ability to generate a set of classification rules efficiently, but they cannot generate rules incrementally when new objects are given. Numerous studies of incremental approaches are not capable to deal with the problems of large database. Therefore, an incremental rule-extraction algorithm is proposed to solve these issues in this study. Using this algorithm, when a new object is added up to an information system, it is unnecessary to re-compute rule sets from the very beginning, which can quickly generate the complete but not repetitive rules. In the case study, the results show that the incremental issues of new data add-in are resolved and a huge computation time is saved.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

The rough set theory, proposed by Pawlak in 1982 [1] can be seen as a new mathematical approach to vagueness [2]. The rough set method does not require additional information about the data; it can work with imprecise values or uncertain data, is capable of discovering important facts hidden in that data, and has the capacity to express them in natural language [3]. In addition, the RS theory is useful today, while a bound of knowledge is surrounded, typically knowledge can be represented in the form of a decision table with rows containing objects and columns containing criteria or attributes. A decision table can be used to derive decision rules through an inductive process. These rules can then be generalized for use in future decision support [4]. The usefulness and effectiveness of the RS approach is shown in data mining, knowledge discovery, pattern recognition, decision analysis, and so on [5,6].

To date, the knowledge discovery literature [7–9] indicates that using RS induct attributes often generates too many rules without focus. These rough set approaches cannot guarantee that the classification of a decision table is credible [10]. Therefore, Tseng [11] proposed the REA (rule-extraction algorithm) to solve the problem. The rule extraction algorithm (REA) was presented for discovering

preference-based rules according to the reducts which contain the maximum of strength index (SI) in the same case. However, the desired reducts are not necessarily unique since several reducts could include the same value of SI. Therefore, an alternative rule can be defined as the rule which holds identical preference to the original decision rule and may be more attractive to a decision-maker than the original one. Thus, Tseng et al. [10] proposed AREA (alternative rule extraction algorithm) to solve the non-complete rules problem.

Moreover, the current algorithms of rough set are capable to generate a set of classification rules efficiently, but they cannot generate rules incrementally when new objects are given. However, the non-incremental approach becomes very costly or even intractable as the number of attributes grows. Alternatively, one can also apply an incremental learning scheme. The essence of incremental learning is to allow the learning process to take place in a continuous and progressive manner rather than a one-shot experience [12]. In practical application, the records of database are often increased dynamically [13]. If new object arrival, it have to compute the whole database again. This procession is due to consume huge computation time and memory space [13].

Most of traditional incremental technique related literatures [14–17] are not capable to deal with the problems of large database. Moreover, to dealing with the new added data set, the traditional methods by re-computing the reduction algorithm and rule-extraction algorithm are often applied [18]. Therefore, Fan et al. [18] proposed an incremental rule-extraction algorithm based

* Corresponding author. Tel.: +886 919 211 312.

E-mail addresses: cchuang@ncnu.edu.tw (C.-C. Huang), bseng@utep.edu (T.-L. Tseng), d96725003@ntu.edu.tw (Y.-N. Fan), a89125339@msn.com (C.-H. Hsu).

on the REA to solve the aforementioned problem. However, alternative rules which are as preferred as the original desired rules might exist since the maximum of SI is not unique. The REA may lead to non-complete rules. Therefore, in this study, an incremental rule-extraction algorithm is proposed based on the AREA [10] to solve the aforementioned problem, named incremental AREA (IAREA). The proposed approach is able to exclude the repetitive rules that maybe generating by original AREA and to avoid the problem of redundant rules.

In summary, considering the insufficient studies in previous literature, the proposed IAREA is able to generate concise and complete alternative decision rules as preferred as the original desired rules. In addition, the proposed incremental structure is capable to address dynamic database problems related to rough set-based rule induction. The IAREA is capable to deal with incremental data solely instead of re-computing the entire dataset when the database is updated. As a result, exceptional computing time and memory space are saved. A case study of CRM is applied to demonstrate validity and efficiency of the proposed method. Since this subject is rarely considered in previous literature, consequently this subject will open a new venue for CRM.

The study is organized as follows: in Section 2, the related literatures are reviewed, while the proposed approach is developed in Section 3. In Section 4, a case study of CRM to demonstrate feasibility of the proposed approach is depicted. Finally, Section 5 concludes this research.

2. Literature review

In this section, the literatures related to the rough set based rule induction, and the related incremental approaches are surveyed.

2.1. Rough set based rule induction

The rough set theory (RST) introduced by Pawlak [1] is a knowledge discovery tool that can be used to help induce logical patterns hidden in massive data. This knowledge can then be presented to the decision-maker as convenient decision rules. Its strength lies in its ability to deal with imperfect data and to classify. It can also be used to analyze the underlying market demand in any industry slated for deregulation [19,20].

In order to gain the meaningful decision rules, two processes are introduced. First, the reduct process of condition attributes determines the superfluous attributes and yields the reduct attribute sets [8]. Second, The rule induction process to determinethe lower and upper approximations, knowledge hidden in information systems may be unraveled and expressed in the form of decision rules. The extracted rules can be used for making predictions in the various domains [21–27].

2.1.1. Attribute reduction process

A reduct is defined as a minimal sufficient subset of a set of attributes which has the same ability to discern concepts when the full set of attributes is used [25,28]. Basically, the reducts represent necessary condition attributes to make a decision. In the attribute reduction process, an algorithm is pre-processing of rule induction, removes redundant information or attributes and selects a attribute subset that has the same discernibility as the original set of attribute [18]. Bazan et al. [29] proposed a dynamic reducts as a tool for extracting laws from decisions tables. They investigated a problem how information about the reduct set changes in a random sampling process of a given decision table could be used to generate these laws. Ślezak [30] used information entropy measure to extend the rough set based notion of a reduct. He introduced the Approximate Entropy Reduction Principle (AERP). It states that any simplification (reduction of attributes) in the decision model,

which approximately preserves its conditional entropy (the measure of inconsistency of defining decision by conditional attributes), should be performed to decrease its prior entropy (the measure of the model's complexity). Chen et al. [31] proposed a reasonable definition of parameterization reduction of soft sets and compare it with the concept of attributes reduction in the rough sets theory. By using this new definition of parameterization reduction, they improved the application of a soft set in a decision making problem. Chen et al. [32] proposed a new approach to attribute reduction of consistent and inconsistent covering decision systems with covering rough sets.

According to the rough set theory, $I = \{U, A\}$ is an information system, where U is a finite set of objects and A is a finite set of attributes. With every attribute $a \in A$, a set of its values Va is associated. Assume $A = C \cup D$, $B \subset C$, where B is a subset of C ; the positive region $POS_B(D) = \{x \in U: [x]_B \subset D\}$ can be defined. The positive region $POS_B(D)$ includes all objects in U which can be with classified into classes of D , in the knowledge B . The degree of dependency between B and D can be defined as $K(B, D) = \text{card}(POS_B(D)) / \text{card}(POS_C(D))$, where card yields the set cardinality. In general, if $K(B, D) = K(C, D)$, and $K(B, D) \neq K(B - \{a\}, D)$, for any $a \in B$ are hold; then B is a reduct of C . Since a reduct (B) preserves the degree of dependency with respect to D and a reduct (B) is a minimal subset, any further removal of condition attributes will change the degree of dependency. The following procedure for determining the reducts and cases is adopted from the Pawlak [25].

Notation:

- U : a finite set of objects;
- A : the abbreviation form of “attribute.”;
- n : the total number of original objects;
- m : the total number of attributes;
- i : object index;
- j : attribute index;
- k : the outcome index.

Input: A decision table I classified into C and D .

Output: The reducts.

Step 1. Initialization: List all objects in I

Step 2. Generate the reducts for each object

for $i = 1$ to n do

for $j = 1$ to m

if $[V_{ij}]_{A_j} \subset [V_{ik}]_{O_{kij}}$

then the reducts for X_i is formed

else for $j = 1$ to m

if $\bigcap_{C-A_j} [V_{ij}]_{A_j} \subset [V_{ik}]_{O_{kij}}$

then the reducts for X_i is formed

else the reducts for X_i is not formed

endfor

endfor

endfor

Step 3. Termination: Stop and output the results.

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات