ELSEVIER

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Physics Procedia 25 (2012) 2025 – 2029

Physics Procedia

2012 International Conference on Solid State Devices and Materials Science

A Data Preprocessing Algorithm for Classification Model Based On Rough Sets

Li Xiang-wei,Qi Yian-fang

1. Department of Computer Engineering, Lanzhou Polytechnic College
2. Key Laboratory of Gansu Advanced Control for Industrial Processes
Lanzhou, China

Abstract

Aimed to solve the limitation of abundant data to constructing classification modeling in data mining, the paper proposed a novel effective preprocessing algorithm based on rough sets. Firstly, we construct the relation Information System using original data sets. Secondly, make use of attribute reduction theory of Rough sets to produce the Core of Information System. Core is the most important and necessary information which cannot reduce in original Information System. So it can get a same effect as original data sets to data analysis, and can construct classification modeling using it. Thirdly, construct indiscernibility matrix using reduced Information System, and finally, get the classification of original data sets. Compared to existing techniques, the developed algorithm enjoy following advantages: (1) avoiding the abundant data in follow-up data processing, and (2) avoiding large amount of computation in whole data mining process. (3) The results become more effective because of introducing the attributes reducing theory of Rough Sets.

Keywords-rough sets; classification; data mining

1. Introduction

Data preprocessing is one of the first and critical step to data mining or data analysis. The results of data preprocessing is directly inputted to mining model and obtained the final results. A good data source can not only increase the accuracy of mining, but also raise the efficiency of algorithm dramatically. In general case, data preprocessing refers to as data cleaning, data integrate, data transition, data reduction, et al., processed before the implementation of data mining algorithm. Whereas, the technique of data mining concern many comprehensive area such as mathematic, computer, statistic, artificial intelligent, computer visual, et al. different application domain need various function of data preprocessing. Luqing proposes a statistical method based on experience, which can transit in attributes using noun [1]. Wang Da-lin, Yu Ge, Bao Yu-bing propose a novel data preprocessing algorithm oriented to Domain knowledge, which introducing the domain knowledge to algorithm for decrease the quantity of data resource[2]. Yang Yang, Liu Feng, Zhang Tian-ge developed a new method to extract feature based on conflict analysis, and effectively eliminate the redundant attributes to classification [3]. Aimed to the data sets feature, Huang Rong-wei, Li Wen-jing discrete the data using the theory of rough sets [4, 5]. Tang Jian-guo, Tan Ming-shu developed a rule extraction method in uncertain environment by using rough sets, concept space and contains degrees [6, 7]. All in all, above mentioned method can increased the efficiency of classification modeling by preprocessing the original data.

Classification is the most fundamental function in data mining and successful used in medical diagnostics, decision analysis, machine learning, information retrieval and approximate reasoning; Whereas, much of above classification techniques cannot get a satisfying results in practice. The main reason is that they all have lower performance during processing because of the affection of abundant data, especially with the rapid increase of original data. To overcome this problem, a novel data analysis tool is introduced, i.e., attributes reduction theory of RS, which has successfully been used in many application domains, such as machine learning, expert system and pattern classification [8, 9].The main advantage of rough sets is that it does not need any preliminary or additional information about original data, such as probability in statistics, or basic probability assignment in Dempster-Shafer theory and grade of membership or the value of possibility in fuzzy sets [10].It can effectively overcome the problem of redundant data and can also preserve the basic and interesting original information.

The rest part of this paper is organized as follows. First, a brief description of underlying theory about this paper is given in Section 2. Section 3 descripts the detailed proposed algorithm based on Rough Sets. Finally, conclusions are drawn in Section 4.

2. Relative theory

2.1 . Foundemental theory of Rough Sets

Let $U \neq \phi$ be a universe of discourse and X be a subset of U . An equivalence relation, R, classifies U into a set of subsets $U/R = \{X1, X2, ..., X_n\}$ in which the following conditions are satisfied:

(1) $X_i \subseteq U, X_i \neq \phi$ For any *i.*

(2) $X_i \cap X_j \neq \phi$ For any *i, j.*

(3) $\bigcup_{i=1,2,...,n}$, $X_i = U$

Any subset X_i , which called a category, class or granule, represents an equivalence class of R. A category in R containing an object $x \in U$ is denoted by $[x]_R$. To a family of equivalence relations $P \subseteq R$, an indiscernibility relation over P is denoted by $IND(P)$, and is defined by equation (1).

$$IND\ (P) = \bigcap_{R \in P} IND\ (R) \tag{1}$$

2.2 Lower and Upper Approximations

The set X can be divided according to the basic sets of R , namely a lower approximation set and upper approximation set. Approximation is used to represent the roughness of the knowledge. Suppose a set $X \subseteq U$ represents a vague concept, then the *R*-lower and *R*-upper approximations of *X* are defined by equation (2) and equation (3).

$$\underline{R}X = \{x \in U : [x]_R \subseteq X\} \tag{2}$$

Equation (4) is the subset of *X*, such that *X* belongs to *X* in *R*, is the lower approximation of *X.*

$$\overline{R}X = \{x \in U : [x]_R \cap X \neq \phi\} \tag{3}$$

Equation(5) is the subsets of all *X* that possibly belong to *X* in *R*, thereby meaning that X may or may not belong to X in R , and the upper approximation $\overline{R}$ contains sets that are possibly included in *X. R*-positive, *R*-negative, and *R*-boundary regions of *X* are defined respectively by equation(4), equation(5) and equation(6).

$$POS_R(X) = \underline{R}X \tag{4}$$

$$NEG_R(X) = U - \overline{R}X \tag{5}$$

$$BNR\ (X) = \overline{R}X - \underline{R}X \tag{6}$$

2.3 Attributes Reduction and Core

In RS theory, an Information Table is used for describing the object of universe, it consists of two dimensions, each row is an object, and each column is an attribute. RS classifies the attributes into two types according to their roles for Information Table: Core attributes and redundant attributes. Here, the minimum condition attribute set can be received, which is called reduction. One Information Table might have several different reductions simultaneously. The intersection of the reductions is the Core of the Information Table and the Core attribute are the important attribute that influences attribute classification.

A subset *B* of a set of attributes *C* is a reduction of *C* with respect to *R* if and only if

(1) $POS_B(R) = POS_C(R)$, *and*

(2) $POS_{B-\{a\}}(R) \neq POS_C(R)$, For any $a \in B$

And, the Core can be defined by equation (7)

$$CORE_C(R) = \{c \in C \mid \forall c \in C, POS_{C-\{c\}}(R) \neq POS_C(R)\} \quad (7)$$

3. The proposed preprocessing algorithm

3.1 Construction of Relational Information System

The majority of research in data mining always concentrated on building the appropriate models for unknown data prediction. Part of the reason, no doubt, is that a prediction task is well defined and can be objectively measured on an independent test sets. Relational Information System is an important and common mathematical tool in data analysis. Many applications can effective solved by constructing the Information System and Information System theory, especially to two dimensional tuple data. If we construct row using objects represented by user, and column using properties represented by P, a classical relation Information System can be denoted as Tab 1.

TABLE I. relation Information System

user	P1	P2	P3	P4	P5	P6	p7	P8	P9
user 1	2	1	2	2	2	1	2	1	2
user 2	1	1	1	1	1	1	1	1	1
user 3	1	2	1	2	1	1	1	2	1
user 4	1	2	1	1	1	1	2	2	1
user 5	1	2	1	1	1	2	1	2	1
user 6	1	1	1	1	1	1	1	1	1

In table I, as previous description, row represents the object or user, and column represents all the properties of each user, the value in table denote the status of each user corresponding to every properties. For the convenience of denotation, the value in table either one or two, we may also represent status using zero and one.

3.2 Generating Reduced Information System

As mentioned in previous section, Core is the most important and necessary properties that cannot be reduced. The main fundamental is indiscernibility relation in Rough Sets. The properties or sub properties in relation Information System can be regard as granularity of knowledge or classification. Accuracy of granularity well embodies the discriminating amplitude of tuple. From table I, we can see obviously that property 1 3 5 9 are identical, and property 2 8 are identical. That is to the classification of user, the property 3 5 9 are redundant property, and property 8 also redundant property. In table I, the columns with red background are all redundant properties. According to the previous description of rough set theory, we can eliminate the redundant property and get core of relation Information System. The reduced relation Information System denoted as table II.

TABLE II. Reduced relation Information System

user	p1	p2	p3	p4	p5
user1	2	1	2	1	2
user2	1	1	1	1	1
user3	1	2	2	1	1
user4	1	2	1	1	2
user5	1	2	1	2	1
user6	1	1	1	1	1

We can see obviously that reduced information system have no redundant properties and the amount of data decreased dramatically. Whereas, to the classification to user1 to user 6, both table I and table II have the same results.

3.3 Construction of Indiscernibility matrix

According to reduced relation information system, and combining the theory of indiscernibility relation, we can easily get an indiscernibility matrix as table III.

TABLE III. Indiscernibility matrix

	user1	user2	user3	user4	user5	user6
user1	0	p1p3p5	p1p2p5	p1p2p3p4	p1p2p3p4p5	p1p3p5
user2		0	p2p3	p2p5	p2p4	0
user3			0	p3p5	p3p4	p2p3
user4				0	p4p5	p2p5
user5					0	p2p4
user6						0

In table III, the row and column are all user, and the value in table denote the different properties to every user, for example, to the user2 and user 6, the value in table equates zero, which denotes that user 1 and user 6 are complete identical in relation Information System, i.e., all the properties of user2 and user6 are identical, so we can subdivide them into same classes. Whereas, to the user2 and user5, the value in table are p1p2p3p4p5, that is to user2 and user5, all the properties are different in Information System, which denote that user2 and user5 are compete different user, we must subdivide them into different classes. To the value in table, the more number of properties, the more different to two users. If value equal to zero, denote the two users complete identical. The similar analysis may use to any other users.

3.4 Getting the cllasification model

From table III, we can classify user into following classes: complete identical users, similar users and complete different users.

①┌ Complete identical users:
{user2, user6}
②┌ Similar users:
{user2, user3}
{user2, user4}
{user2, user5}
{user3, user4}
{user3, user5}
{user3, user6}
{user4, user5}
{user4, user6}
{user5, user6}
③┌ Complete different users:
{user1, user5}

From previous description, we can classify original relation information system user into three classes: complete identical users: user2 and user6, similar users: and complete different users: user2 and user5. In practice, the results can help us making effective decision for further analysis or application such as authoritative user mining or personalization servers recommend.

4. Results and discussion

Classification algorithm, usually called supervised learning in data mining, is one of the most important and fundamental function in data analysis. The final goal of classification is to build a set of models that can correctly predict the class of the different objects, the input to this methods is a set of objects(i.e., training data), the classes which these objects belong to (i.e., dependent variables), and a set of variables describing different characteristics of the objects(i.e., independent variables). Once such a predictive model is built, it can be used to predict the class of the objects for which class information is not known a priori. The key principle of classification is to extract classification model according to training sets, and then classify unknown data using model. While in practice, many existing method have a low efficiency since without any effective preprocessing techniques or mechanism to redundant data before constructing classification modeling. Redundant attribute in data set can lead to low precision and low interpreting ability of data mining and virtually determine poor performance of algorithm. On the basis of characteristic analysis to redundant data, introduced the novel mathematical theory, the paper developed an effective data preprocessing algorithm for classification model. The proposed algorithm can widely used in data analysis or any other application domain. A concrete instance verified the feasibility.

References

[1] Luqing, “a preprocessing method for nominal attributes in classfication and prediction problems,” computer engineering, 2004,vol. 30, no.3, pp. 92–94.

[2] Wang Da-lin, Yu Ge, Bao Yu-bing, “a representation about domain knowledge for reprocesses of data mining,” mini- micro systems, 2003,vol. 24, no.5, pp. 863–83.

[3] Yang Yang, Liu Feng, Zhang Tian-ge, “preprocessing data for classifier,” computer engineering, 1998,vol. 24, no.4, pp. 33–39.

[4] Huang Rong-wei, Li Wen-jing, “data preprocessing based on rough set theory,”journal of guangxi teachers education university, 2006,vol. 23, no.4, pp. 87–95.

[5] Tang Jian-guo, Tan Ming-shu, “on finding core and reduction in rough set theory,”control and decision, 2003,vol. 18, no.4, pp. 449–457.

[6] Di Kai-chang Li De-ren, Li De-yi, “rough set theory and its application in attribute analysis and kenowledge discovery in gis,”journal of wuhan technical university of surveying and mapping, 1990,vol. 24, no.1, pp. 1–10.

[7] Song Xiao-xue, “rough sets theory and its application,”journal of xianyang normal university, 2005,vol. 20, no.2, pp. 30–35

[8] Yiyu Yao, “notes on rough set approximations and associated measures,” journal of zhejiang ocean university(natural science),2010, vil. 29, no. 5, pp. 399-410.

[9] Yiyu Yao, and Zhao, Y., “attribute reduction in decision-theoretic rough set models,” informaion sciences, 2008, vol. 178, no. 17, pp.3356-3373.

[10] S Kotsiantis, D. Kanellopoulos, P.Pintelas, “data preprocessing for supervised leaning,” internal journal of computer science, 2006, vol.1, no.2, pp. 111-117