TELKOMNIKA, Vol. 11, No. 7, July 2013, pp. 3597 ~ 3603
e-ISSN: 2087-278X

Two-step Classification Algorithm Based on Decision-Theoretic Rough Set Theory

Jun Wang*[1], Yulong Xu[2], Weidong Yu[1]
[1]Textiles College, Donghua University, Shanghai 201620, PR China
[2]Institute of Information and Technology, Henan University of Traditional Chinese Medicine, Zheng Zhou 450003, PR China
Corresponding author, e-mail: cnwjun@Hotmail.com, flyxyl@126.com, wdyu@dhu.edu.cn

Abstract

This paper introduces rough set theory and decision-theoretic rough set theory. Then based on the latter, a two-step classification algorithm is proposed. Compared with primitive DTRST algorithms, our method decreases the range of negative domain and employs a two-steps strategy in classification. New samples and unknown samples can be estimated whether it belongs to the negative domain when they are found. Then, fewer wrong samples will be classified in negative domain. Therefore, error rate and loss of classification is lowered. Compared with traditional information filtering methods, such as Naive Bayes algorithm and primitive DTRST algorithm, the proposed method can gain high accuracy and low loss.

Key words: *rough set; decision-theoretic rough set; classification*

Copyright © 2013 Universitas Ahmad Dahlan. All rights reserved.

1. Introduction

Rough Set Theory proposed by Professor Z. Pawlak in 1982 [1-2], has been widely used in multi-fields such as the machine learning, data mining and so on since it can be used to deal with and analyses kinds of inaccurate, inconsistent and incomplete information and reveal the potential rule [3-7]. Besides, Rough Set has been alppied in others real problem includes extracting decision rules [8] and 3D technology [9].

Decision-theoretic rough set theory (DTRST) proposed by Yao and Wong [10-12], was enhanced to process inaccurate information through inheriting all of the basic properties of the original model of RS and introducing the Bayesian decision theory. It is proved that DTRST is more general than RST, which can translate into different type of rough set model, and sort the information. According to the DTRSR, an information filtering algorithm was proposed, which can reduce the loss in assortment, and be better to Bayesian algorithm. However, there is still being large loss in new sample classification because of the expanding scope of the negative domain. In this article, a two-step information filtering algorithm is proposed. Under the premise of correct rate, it can reduce loss as much as possible in the classification. Especially, the loss would be reduced to minimum used by this two-step algorithm in the new or unknown samples. Compared to the general algorithm, this one is better, and the correctness of which has been proved by the simulation experiment.

The structure of this article is as follows: The rough set and decision-making related concepts rough sets are introduced in Section 2. A two-step classification algorithm is proposed based on decision rough set in Section 3.Simulation experiment and analysis of results is shown in Section 4. The conclusion is in the last Section.

2. Related Concepts

For convenience, here, some basic concepts of rough set are made a brief description at first.

Definition 1 (Decision Table) In a decision table, U called as domain of discourse means a set of object. $A = C \cup D$ is called as a collection of properties. C and D are respectively called condition attributes and decision attributes. V means the collection of

Received January 15, 2013; Revised April 5, 2013; Accepted April 15, 2013

attribute values. $f:U\times A->V$ is a function of information, which assigns the attribute values for each object x in U.

Definition 2 (Indiscernibility Relation) In a given decision table as $S=<U,A=C\cup D,V,f>$, for each subset of attributes as $B\subseteq A$, we define an indiscernibility relation as $IND(B)$, which means $IDN(B)=\{(x,y)\,|\,(x,y)\in U\times U,_{\forall b\in B}(b(x)=b(y))\}$. Obviously, the indiscernibility relation is an equivalence relation.

Definition 3 (Upper/lower approximation set) In a given decision table as $S=<U,A=C\cup D,V,f>$, for each subset of attributes as $X\subseteq U$ an indiscernibility relation as B, the upper//lower approximation set of X is defined as:

$$B^{-}(X)=\cup\{Y_i\,|\,Y_i\in U/IND(B)\wedge Y_i\cap X\neq\varnothing\}$$

$$B\_(X)=\cup\{Y_i\,|\,Y_i\in U/IND(B)\wedge Y_i\subseteq X\}.$$

Definition 4 (Rough membership function) In rough set theory, on the basis of the existing knowledge and indiscernibility relation as B, the uncertainty about that element x belongs or not to the set X, can be presented by rough membership function. It is defined as follows: μ_X^B : U→[0,1] in addition $\mu_X^B(x)=\frac{|[x]_B\cap X|}{|[x]_B|}$;

Rough membership function can be understood as the conditional probability estimates $\Pr(x\in X\,|\,u)$ based on frequency. On the given knowledge U, the object x belongs to the conditional probability of set X. Where U is the characteristics of the object xfor attribute set B.

Definition 5 (Region and Rough membership function) Rough membership function can be understood as the conditional probability estimates $\Pr(x\in X\,|\,u)$ based on frequency. On the given knowledgeU, the object x belongs to the conditional probability of set X. Where U is the characteristics of the object xfor attribute set B. Through rough membership function, the positive region POS (X), the negative domain NEG (X) and the boundary region BND (X) of set X of could be defined respectively as follows:

POS(X) = $\underline{B}_{\pi}X$ =$\{x\,|\,\mu_X^B(x)\geq\pi\}$

NEG(X) = U-$\overline{B}_{\pi}X$ = U-$\{x\,|\,\mu_X^B(x)>1-\pi\}$

BND(X) = $\overline{B}_{\pi}X-\underline{B}_{\pi}X$

Among these, $\pi\in(\frac{1}{2},1]$ which represents the roughness or accuracy,can be regarded as the threshold when dividing object.

Here, some related concepts used in this articlefor decision-making rough set theory are introduced.

Definition 6 (Decision-Theoretic Rough Set) DTRST [10-12] proposed by Yao and Wong, was enhanced to process inaccurate information through inheriting all of the basic properties of the original model of RS. In this theory, the state set Ω = {X, ¬ X} is as whether an element belonging to the set X.Action set A_x={a_1, a_2, a_3}means the action that determining whether the current object X belongs to the POS (X), NEG (X) and the BND (X).

a_1、a_2、a_3 means the determine current objects x∈POS(X) 、x∈NEG(X) 、x∈BND(X)

Definition 7 (Decision loss value) [10]**)** The set λ(a_i|x∈X) is the loss caused bythe executed action a_i, if the condition is x∈X. Therefore, the estimated loss value EL (a_i|x) for the three different activities are as follows:

$$EL(a_1|x)=\lambda_{11}P(X,x)+\lambda_{12}P(\neg X,x)$$

$$EL(a_2|x)=\lambda_{21}P(X,x)+\lambda_{22}P(\neg X,x)$$

$$EL(a_3|x) = \lambda_{31}P(X,x) + \lambda_{32}P(\neg X,x)$$

Among these, P(X, x) and P(¬X, x) indicate the probability that x belongs to X and x belongs to ¬X, $\lambda_{i1} = \lambda$ (a_i |x∈X), $\lambda_{i2} = \lambda$ (a_i |x∈¬X), i=1, 2, 3。

Definition 8 (Bayesian decision rule [10-12]**)** According to the Bayesian decision-making process, minimum risk decision rule can be deduced as $RUL_{P,N,B}$, RUL_P, RUL_N, RUL_B:

RUL_P : if EL(a_1|x)≤EL(a_2|x) and EL(a_1|x)≤EL(a_3|x), then x∈POS(X).

RUL_N : if EL(a_2|x)≤EL(a_1|x) and EL(a_2|x)≤EL(a_3|x), then x∈NEG(X);

RUL_B : if EL(a_3|x)≤EL(a_1|x) and EL(a_3|x)≤EL(a_2|x), then x∈BND(X)。

Because of P(X, x) + P(¬X, x) =1, The above decision rule can be simplified to the form of P (X, x) which only contains the probability. Therefore, we can divide the belonging region for object x through the P(X, x) and a given loss functions λ_{ij} (i=1,2,3; j=1,2).

When the condition is $\lambda_{11} \leq \lambda_{31} < \lambda_{21}$ 且 $\lambda_{22} \leq \lambda_{32} < \lambda_{12}$,for the object x∈X, the loss of dividing x into the positiveregion POS (X) islessthan into the boundaryregion BND (X), furthermore the loss of both above is strictly smaller than the loss of diving x into the negative region NEG (X) .Conversely, the object not belonging to X be divided into X will introduce the reverse order. To these type of loss function, the minimum risk decision rule $RUL_{P,N,B}$, RUL_P, RUL_N, RUL_B, could be written as:

RUL_P : if P(X , x)≥β and P(X, x)≥γ, then x∈POS(X);

RUL_N : if P(X , x)≤γ and P(X , x)≤δ, then x∈NEG (X);

RUL_B : if δ ≤P(X , x) and P(X , x)≤β, then x∈BND(X)。

Among these:

$$\beta = \frac{\lambda_{12} - \lambda_{32}}{(\lambda_{31} - \lambda_{11}) + (\lambda_{12} - \lambda_{32})}$$

$$\gamma = \frac{\lambda_{12} - \lambda_{22}}{(\lambda_{21} - \lambda_{11}) + (\lambda_{12} - \lambda_{22})}$$

$$\delta = \frac{\lambda_{32} - \lambda_{22}}{(\lambda_{21} - \lambda_{31}) + (\lambda_{32} - \lambda_{22})}$$

From the condition $\lambda_{11} \leq \lambda_{31}$, λ_{21}, $\lambda_{22} \leq \lambda_{32}$, λ_{12}, we can know $\beta \in (0,1)$, $\gamma \in (0,1)$, $\delta \in [0,1]$. In this case, the decision rule $RUL_{P,N,B}$ only depends on the parameter β, γ and δ, which could be directly calculated by given value λ_i from the user.

As if $\delta \leq \beta$, $\delta \leq \gamma \leq \beta$, according to the $RUL_{P,N,B}$, the positiveregion, negative region and boundaryregion can be decided by δ and β. As if $\beta < \delta$, and then $\beta < \gamma < \delta$. According to $RUL_{P,N,B}$, we know the boundary region is null, so the positiveregion and negative region can be decided by γ.

In order to distinguish the three regions, we make $\delta < \beta$ and obtain $\delta < \gamma < \beta$.

Further, the risk of x∈NEG (X)and x∈BND(X) is determined the same, then it is determined x∈BND(X). If the risk of x∈POS(X)and x∈BND(X) is determined the same, then it is determined x∈POS(X). Under these assumptions, the decision-making rules can be simplified as:

RUL_P : if P(X, x)≥β and $\delta \le \gamma \le \beta$, then x∈POS(X);

RUL_N : if P(X, x)<δ , then x∈NEG(X);

RUL_B : if δ ≤P(X,x)<β , then x∈BND(X)。

3. Two-step Classification Algorithm Based on Decision-Theoretic Rough Set

In order to achieve the information classification, we consider a loss of function, that the loss is 0 if a virtual object x belonging to X was classified into the positive region POS (X) (i.e., x is actually positive object, and also is divided into positive objects, there is no loss). Conversely, the loss is1 if a virtual object x belonging to X was classified into t negative region NEG (X). The loss of divided into boundary region is a value between 0 and 1. In view of the above set:

$$\lambda_{11} = 0,\quad \lambda_{12} = 1$$

$$\lambda_{21} = 1,\quad \lambda_{22} = 0$$

$$\mathbf{0} \le \lambda_{31} < 1,\quad 0 \le \lambda_{32} < 1$$

$\lambda_{12} = 1$ and $\lambda_{21} = 1$ is design loss value, and can also be the other value, but the relationship between them must met $\lambda_{21} \ge \lambda_{12} \ge 1$, meaning that the loss value of object belonging to positive region was divided into negative region is greater than the loss from the converse process.According to the above values, the estimated loss function can be simplified into the following form:

$$\beta = \frac{1-\lambda_{32}}{\lambda_{31}+1-\lambda_{32}} \quad \delta = \frac{\lambda_{32}}{1-\lambda_{31}+\lambda_{32}}$$

This, if we are able to estimate the value of P (X, x) and the loss value λ_{31} and λ_{32} , we could classify the object. On the base of analysis about P(X,x) in the reference 5, a new algorithm is proposed in this article. We make that

$$P(X,x) = \sqrt{\frac{\sum_{1}^{n} Support_x^2}{N_x}}$$

Where in, N_x represents the total number of object x matching rule in X, $Support_x$ represents the information which interest to the user, its value is accuracy(a→b) [11]:

$$accuracy(a \to b) = \frac{support(a \cdot b)}{support(a)}$$

From the evidence of a to conclusion b, the rules could be proved to credibility, and estimates the conditional probability $P(b \mid a)$ based on the frequency, hence, the accuracy(a→ b) is the same as the value of rough membership function μ_B^X , in which, membership function is applied to a matching object x. P(X, x) is the arithmetic mean of all rules matches Credits, we know that the square mean value is greater than or equal to the arithmetic mean, namely:

$$\sqrt{\frac{\sum_{1}^{n} Support_x^2}{N_x}} \ge \frac{\sum_{1}^{n} Support_x}{N_x}$$

In this article, the square mean is credibility close to the true meaning of the classification, which can reduce the loss in calculation. In summary, the new algorithm proposed is as follows:

Input: decision rules table RUL, test set TE.
Output: three categories about LIKE, UNLIKE, MAYBE
Step 1: each record x in TE is matching with interesting decision-making rules in table RUL.
Step 2: If the record matches the number is zero, turn to Step3, otherwise turn to Step4.
Step 3: match the record with the uninteresting decision-making rules in table RUL, if the number of matches is 0, then this recording was classified as MAYBE, otherwise the recording was classified as Unlike, then turn Step1.
Step 4: If the record matches the number is N, then its Average credibility is

$$P(X,x)=\sqrt{\frac{\sum_{1}^{n} Support_x^2}{N_x}}.$$

Step5: If $P(X,x) \geq \beta$, then judge $x \in LIKE$, If $\delta \leq P(X,x) < \beta$ then judge $x \in MAYBE$, If $0 < P(X,x) < \delta$ then judge $x \in UNLIKE$.

When the general DTRST algorithm is used to classify, the match number of record x and interesting rule is 0 ($P(X,x)$ =0),then x is judged as $UNLIKE$, which expanded the area of $UNLIKE$. Which could make a new interesting record x to be judged to $UNLIKE$. It causes a great loss. In our new algorithm, such a miscarriage of justice could be avoided, and loss could be reduced during to the two step strategy in the classification.

4. Simulation Test

Simulation experiment is conducted using VC + +6.0 development tools in the windows environment. Computer Configuration: CPU Intel Pentium 2.4G; memory 1G; OS Windows XP. Platform is developed by the Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications RIDAS (Rough Set Intelligent Data Analysis System), which is integrated more than 30 classic algorithmsabout Rough Set .

The experimental data is from the UCI machine learning database (http://www.ics.uci.edu/mlearn/MLRepository.html). Four-fifths Experimental data objects act as TR (3681 objects), one-fifths of the data objectsact as TE (920).

In order to verify, the experiment is divided into three parts, and we compared the new improved algorithm with Naive Bayes algorithm and the original decision rough set algorithm. We make the experiment parameters $\lambda_{31}=\lambda_{32}$=0.2, and then β=0.8, δ=0.2, the specific results are as follows.

Experiment 1: Naive Bayes algorithm, TR (4/5, 3681 objects).

Table 1. Class results based on Naive Bayes algorithm

Actually	Predict	
	Like	Unlike
Like	544	28
Unlike	40	308

Experiment 2: General DTRST algorithm, TR (4/5, 3681 objects).

Table 2. Class results Based on the original DTRST algorithm

Actually	Predict		
	Like	Unlike	Maybe
Like	510	6	56
Unlike	62	219	71

Experiment 3: Two steps DTRST algorithm proposed in this article, TR (4/5, 3681 objects).

Table 3. Class results Based on Two steps DTRST algorithm

Actually	Predict		
	Like	Unlike	Maybe
Like	510	5	57
Unlike	62	212	78

In the experimental results, it is shown that to identify 572 interested mails. For the correct ones, the firstalgorithm identifies the maximum number, the latter two followed. However, for the errors, the first up second followed by, and the third at least. Where it is, the loss of dividing one interested mailas spam one is much larger than that to divide it as suspicious one. Though the latter two algorithms to identify the correct number is less than the first one, the risk was also reduced.The total loss was reduced whenthe important and interested mails determined to be spam ones as little as possible.

After careful analysis the results of the latter two, to identify the number of error emails, the proposed algorithm is less than the second. To the suspicious ones, the proposed algorithm is greater than the second. The reason of these conditions is from the two-step identification strategy. During to the two-step strategy, when interested mail not match with all the rules first appears, and fail to match with the original DTRST algorithm identification process, it would be identified as spam.However, the algorithm proposed in this paper is a two-step strategy, after failing to match with the interested rules, it would be matched with the uninterested rules before giving results. This step-by-step strategy has narrowed the scope of the spam, and reduced the mis-division the interested emails into spam as little as possible. This method could increase the number of suspicious mail, but as a whole, this algorithm would reduce the risk and loss more.

After comparing the number of interested emails, the results are the same and not improve compared to the original algorithm. We know that the loss of dividing the sample belonging to positive region into negative region is much greater than its inverse process. The proposed algorithm is a further narrowing the range of negative region. Although the number of interested mails does not improve, the total loss could be reduced. The rules of two algorithms used in classification are the same, and the same division for the interested mails. To further improve the classification accuracy of mails and optimize the rules by adding incremental learning technology, which can enhance the new samples and unknown sample processing capabilities, which will also be our next task.

In summary, the two-step classification algorithm proposed in this article based on the original DTRST algorithm is an improved algorithm on the decision-making rough set theory.

5. Conclusion

Decision-rough set is the general expansion of rough set, which enhance the processing of uncertain information and reduce the loss of classification information. The algorithm proposed in this article, is an improved algorithm on the decision-making rough set theory. The experimental results show that this algorithm could further improve accuracy, reduce risk and loss in classification compared to the original decisions. In future work, we will combine the incremental learning technology in the information classification to make the algorithm generates independent study rules when encountering new messages or unknown information, so that the loss would be reduced at the same timeto further improve the accuracy of the classification.

References

[1] Pawlak. Rough Set. *International Journal Of Computer and Information Sciences.* 1982; 11(5): 341-356.
[2] Z Pawlak. Rough Sets: Theory and Its Application to Data Analysis. *Cybernetics and Systems*. 1998; 29(7): 661-688.

[3] A Czyzewski. Automatic Identification of Sound Source Position Employing Neural Networks and Rough Sets. *Pattern Recognition Letters*. 2003; 24(6): 921-933.
[4] Liu SH, Sheng QJ, Wu B. Rough Set Theory and Efficient Algorithm.*Chinese Journal of Computers*. 2003; 5(26): 524-529. (in Chinese)
[5] Wang GY. Extension of rough set under incomplete information systems. *Journal of Computer Research and Development* (in Chinese). 2002; 39(10): 1238-1243.
[6] Yao YY. Probabilistic approaches to rough sets, *Expert Systems*. 2003; 20(5); 287-297.
[7] Meng Y, Luo K, Liu JH, Jiang F. SA Rough Sets Kmeans resource dynamic allocation strategy Based on Cloud Computing Environment, *TELKOMNIKA Indonesian Journal of Electrical Engineering,* 2012;10(6): 1485-1489.
[8] Hossam M. An Integrated Methodology of Rough Set Theory and Grey System for Extracting Decision Rules. *International Journal of Advances in Applied Sciences*. 2013; 2(1), in pressing
[9] Guoyin Wang, Yan Wang, 3DM: Domain-oriented Data-driven Data Mining, Fundamenta Informaticae. 2009; 90(4): 395-426.
[10] Yao YY. Decision-Theoretic Rough Set Models, Rough Sets and Knowledge Technology, 2007; LNAI 4481. 1-12.
[11] Zhao WQ, Zhu YL, Gao W. Information Filtering Model Based on Decision Making Rough Set Theory. *Computer Engineering and Applications* (in Chinese). 2007; 43; 185-194.
[12] Yao YY, Wong SK. A decision theoretic framework for approximating concepts. *International Journal of Man-machine Studies*. 1992; 37(6); 793-809.