Data Mining Based on the Rough Analysis and the Application in the Telecommunication Network Quality Evaluation

Yujuan Guan*, Wenyu Zhang
College of Economics and Management, Xi’an University of Post & Telecommunications, Xi’an, 710061, China
*Corresponding author, e-mail: shafu6382964@yeah.net

Abstract
The precision of the classification rule is decided by the construction of the classification algorithm. In this paper, after the introducing of the concepts and the attribute-reduction algorithms of the basic rough set, a data-mining algorithm based on the rough imped approximation measurement is demonstrated. At first, the finding theory of the classification rule is introduced. After that, according to the importance-measurement of the conditional attributes in the decision table, one kind of new attributes reduction method in the field of data mining, which is based on the improved rough imped approximation measurement, is explained. Lastly, the method is applied in the telecommunication network quality evaluation, and the rationality and feasibility of algorithm are examined in this example.

Keywords: data mining, rough set, classification approximation, telecommunication network quality evaluation

1. Introduction
Data Mining is concern with mining the valuable data from the massive data in the database to assist the decision-making. Presently, data mining is one of the research directions in the field of Information Decision and Artificial Intelligence. The Rough set method is a new information processing method [1, 2], which is based on the classification of the equivalence relation. One of the characteristics of this method is, it is not needed to give the quantity description of some attributes in advance, rather finding out the internal rules of the problem directly from the description set of this problem [3, 4]. It has great capability in analyzing the information system, which is incomplete, inaccuracy and includes noise.

This Rough set is attended in the world because it has been successfully utilized in the field of Data Mining and Knowledge Discovering in database in recent years. There are three classification of attribute reduction based on rough set: (1) Attribute reduction algorithm based on discernibility matrix, reduction algorithm based on heuristic greedy, (2) Attribute reduction algorithm based on information indicated, and (3) Attribute reduction algorithm based on discernibility matrix was originally proposed by Skowron. In this algorithm, the problem of knowledge reduction of the decision table of information systems was converted to the problem of the reduction of discernibility matrix. Simple, easy to understand and achieve in practical application is the feature of the algorithm, while the disadvantage is that the algorithm reduction is inefficient and is a serious waste of space, not suitable for mass data.

To address the above problem, various improved modifications of the difference matrix algorithm was got by reducing the time complexity of difference matrix reduction and many other ways by many scholars, making it a more efficient reduction algorithm. A property selection criteria was designed by greedy heuristic attribute reduction algorithm as the basis for selection of the current best attribute. Using heuristic ideas to reduce properties not only has the advantages in terms of time efficiency, and the final result is usually the best reduction or suboptimal, in fact, our needs in practical application can be satisfied by this result, so the heuristic algorithm is generally preferred as the attribute reduction.

Rough set theory is generally indiscernible relationship-based, by introducing the approximation set and the lower approximation set, defined on the set operations, this is often
called the algebraic point of Rough set theory. However, some scholars studied the Rough theory and put forward an information theory idea of Rough theory, which based on the theory of information. For this reason, it makes possible to studying the attributes reduction of rough set, which based on the information theory of Rough theory. Attribute reduction algorithm based on information indicated, that is based on information theory. One of the most representative reduction algorithm are with the mutual information of condition attributes and decision attributes and with the conditional entropy of decision attribute set to conditional attributes, which as a measure tool of attribute significance. In this thesis, the second attribute reduction method was adopted, and an improved rough approximating approximation measurement in date mining method was proposed. Thus an attribute reduction is effectively and applying the method to the telecommunications network quality assessment.

2. Algorithm optimization

2.1. Conception

\[
\begin{align*}
R_2(X) &= \left\{ a : u \in U \text{and} B(u) \text{is the subset of } X \right\} \\
R_2'(X) &= \left\{ u : u \in U \text{and} B(u) \cap X \neq \emptyset \right\}
\end{align*}
\]

Definition 1.1 Let \( S = (U, A, \{V_a\}, \alpha) \) be knowledge expression system. Where, \( U \) is not null limited set calling an universe and \( A \) is not null limited set calling attribute sets; \( V_a \) is the value domain of attribute \( a \in A \), and the expression of \( \alpha : U \rightarrow V_a \) is a single image which can correspond any element in \( U \) with a certain alone value in \( V_a \). If \( A \) is composed by condition attribute set \( C \) and decision attribute set \( D \) and \( C, D \) meet the demands of \( C \cap D = \emptyset \) and \( C \cup D = A \), \( S \) is called decision system. In briefly, let \( (U, C \cup \{d\}) \) be decision system, namely decision attribute set only concludes one element \( d \).

\[
\text{ind}(B, \{d\}) = \{(x, y) \in U \times U : d(x) = d(y)\}
\]

\[
\text{ind}(B, \{d\}) = \{(x, y) \in U \times U : d(x) = d(y)\}
\]

Definition 1.2 In the decision system \( S = (U, A, \{V_a\}, \alpha) \), when \( B \) denote a set whose elements belong to set \( C \) and the binary relation is called the indiscernibility relation of set \( S \). Among this system, \( x \) and \( y \) are the elements of set \( U \), also \( d(x), d(y), a(x), a(y) \) denote the attribute values of \( x, y \) in attribute \( d \) and attribute \( a \).

Definition 1.3 The knowledge expression system is defined as \( S = (U, A) \), where \( U \) is the universal set and \( A \) is attribute set. The set \( B \) is the subset of \( A \) and \( \text{ind}(B) \) is the equivalence relation of \( U \times U \). \( B(u) \) which is called \( B \)-base set is a equivalence class concluding element \( u \) coming from equivalence relation \( \text{ind}(B) \). An equivalence class sets can be obtained by classifying \( U \) with attribute set \( B \). An upper approximation set of \( X \) in \( B \) and a lower approximation set of \( X \) in \( B \) are separately defined as:

Definition 1.4 Attributes Dependency: The importance of conditional attributes set \( C \) constrained by decision attribute set \( D \) has been determined by dependencies between attributes set \( D \) and set \( C \). That decision attribute set \( D \) on condition attributes set \( C \)'s degree of dependence can be indicated using the following definition:

\[
\gamma_C(D) = \frac{|\text{POS}_C(D)|}{|U|}
\]

\[
\text{POS}_C(D), \text{ called a positive region of the partition } U/\text{IND}(D) \text{ with respect to attributes set } C.
\]

Definition 1.5 Attributes Significance: Different attributes can play different roles in the dependency relationship between condition attributes and decision attribute. The significance of that attribute \( a \) to join \( C \) by indiscernibility relation \( U/\text{IND}(D) \) can be defined as:
The data-mining model based on the rough set theory for the decision system depends on the reducing of the decision system. In order to form the classification rules, there are two kinds of reducing method, one is the attribute reducing, which means eliminating the unnecessary line in the decision table, and another is the attribute value reducing, which means eliminating the unnecessary attribute value in the decision table. Rampure and Tiwari (2014) and Liu et al. (2012) the decision table before the reducing has the same functions as the reduced decision table. But if both of them make the same decision, the reduced decision table needs fewer conditions. That means we can achieve the same objective by the simpler way. Normally, there are many ways to reduce the decision table. So the problem is which one is the best one? This electing process should depend on the best rules that relate to the attributes [5]. If there is some kind of cost function for the attributes, the rule could be that the attributes that have the least cost should be eliminated. In the case of no cost function, the only information resource for the best reducing is the decision system.

2.2. The Principle for Finding the Classification Rules

Forming classification rules in the decision system is the key-step in the data mining [6-10]. It depends on the conclude-depend relation between the data. The conclude-depend relation is defined as the function relation that is consistent with the current database but not clarified in the model of this database. In another words, the conclude-depend relation means when the description of the data entity in the attributes set C is illuminated, the value of the attribute D of this data entity can be determined. At this time we say there is the conclude-depend relation C→D between the attribute set C and attribute D, namely, D is depending on C.

On the reason of the border area in the data set, there are two sorts of conclude-depend relations, the strong conclude-depend relation and the weak conclude-depend relation. The strong conclude-depend relation means the description of the data in the condition attribute set C can determine the only value of the decision attribute. When there is no border area in the data set, the relationship between the condition attribute and the decision attribute is the strong depend relation; the weak conclude-depend relation means the description of the data in the condition attribute set C can determine the value of the decision attribute with the certain probability. When there exist the border area in the data set, the relationship between the condition attribute and the decision attribute is the weak depend relation.

Calculating the least conclude-depend relation means reducing the attributes by eliminating the redundant attributes in the condition attribute set [11-15].

There are two sorts of redundant attribute. One is not relating to the decision attribute and has no conclude-depend relation with other condition attribute, called the first sort redundant attribute. Another is not relating to the decision attribute also but has the conclude-depend relation with other condition attribute, called the 2nd sort redundant attribute. For one existing decision system, finding all of the attribute reductions and the attribute value reductions is one complete NP problem. In the real life, normally we just concern the least reduction; there is no need to find all of the reductions. Because the core is concluded in all of reductions, starting with the core is the efficient way to calculate the least reduction of the table.

In this paper, using the concept of the impend approximation measurement, we start with the core of the decision table and according to the measurement of the importance of the attribute add the most important attributes into the core until meeting the requirement, then we can get one reduction of the decision table. In this paper, from the impend approximation measurement of every condition attribute set in the decision table to the decision classification, one improved rule for measuring the importance of the attribute is described [16-21].

2.3. Improved Impend Approximation Measurement Model

The dependent variables of presumable range is deleted by the Knowledge reduction technology, that is, The presumable range of each relationship is invariant in the operation process, which means the Indiscernibility set of probability estimation can be fixed up. Notice that the deterministic case was contained by this definition, if SI is identically equal to 1, that is, indiscernibility set of a rule corresponds to the positive domain of a decision classification. From the angle of conditional probability, reduction ideas were put forward for attributes reduction in this paper.
The decision system is made as: \( S = \langle U, C \cup \{d\}, V, f \rangle \), the value of approximate space measure \( \gamma \) is given, if \( \text{POS}_{C-A}(\gamma(Y)) = P_c(Y) \) is right, the attribute \( a \) of attribute set \( C \) is dependent in view of the decision set \( \{d\} \), otherwise, it’s independent. For the condition attribute set \( K \subset B \subseteq C \), if \( \text{POS}_K(Y) = \text{POS}_K(\gamma(Y)) \) existed, the condition attribute set \( B \) of decision system \( S \) is called dependent set in view of the decision attribute \( \{d\} \), otherwise, it’s called independent set \( A \) reduction \( C' \) of attribute \( C \) in condition attributes is a maximal independent subset in the view of decision attribute \( \{d\} \). The process of probability reduction to find condition attributes were similar to the traditional rough set attribute reduction process, that \( \gamma \) positive domain of set \( Y \) is \( \{Y_{POS}, Y_{POS} \} \), then attribute \( a \) can be marked as a condition attributes and be removed, do the same with the other condition attributes, the remaining condition attributes in the condition attribute set is a reduction for property. Based on the core of condition attributes, an attribute set reduction approach was proposed by making use of the parameters of the approximation accuracy of condition attributes in this paper.

It is an approximation classification question which the knowledge is obtained from the set and classification in some objects in knowledge. In order to describe the uncertainty of approximation classification, the classification approximation can be defined. The definition of the classification approximation is the simple extension of the set approximation. In set \( L = \{Y_1, Y_2, \ldots, Y_n\} \) where \( U = \bigcup_{i=1}^{n} Y_i \), \( L \) is the set groups of \( n \) classification in set \( U \) and the classification is carried out based on knowledge.

Namely let \( Y_i = \{y_1, y_2, \ldots, y_k\}, i = 1 \sim n \) be a classification, an upper and lower approximation set of \( R \) in \( Y \) are defined as:

\[
R_+(L) = \{R_+(Y_1), R_+(Y_2), \ldots, R_+(Y_n)\} \quad \text{and} \quad R_-(L) = \{R_-(Y_1), R_-(Y_2), \ldots, R_-(Y_n)\}
\]

Definition 3.1 let \( S = \langle U, A \cup \{d\} \rangle \) be a decision system, \( P \subseteq Q \) is subset, \( L \) is the classification of \( \{Y_1, Y_2, \ldots, Y_k\}, Y \subseteq U \) decided by decision attribute, then the impend approximation measurement of set \( Y \) about attribute \( P \) or calling rough degree is defined as:

\[
\alpha_p(Y) = \frac{\text{card}(\overline{P}(Y))}{\text{card}(P(Y))}
\] (4)

Where \( \text{card}(Y) \) expresses the card of the set \( Y \).

Define the impend approximation measurement of decision classification about attribute \( P \) or approximation classification quality as:

\[
\gamma_p(L) = \sum_{i=1}^{k} \frac{\text{card}(P(Y_i))}{\text{card}(U)}
\] (5)

Define the rough impend approximation measurement of decision classification \( L \) about attribute \( P \) or approximation classification approximation measure as:

\[
\alpha_p(L) = 1 - \left( \frac{\sum_{i=1}^{k} \text{card}(\overline{P}(Y_i))}{\sum_{i=1}^{k} \text{card}(P(Y_i))} \right) / \left( \sum_{i=1}^{k} \text{card}(\overline{P}(Y_i)) \right)
\] (6)

The impend approximation measurement (approximation classification quality) expresses percentage of precisely classifying into \( Y \) with knowledge \( P \), and the rough impend approximation measurement (approximation classification approximation quality) describes the concise decision percentage in the possible decision in the time of classifying the objects using knowledge \( P \).

Define the classification quality measure of decision classification \( L \) about attribute \( P \) as:

\[
\phi_p(L) = \frac{k}{\text{card}(U)} \cdot \gamma_p(L)
\] (7)
Define the relative classification of decision classification $L$ about knowledge $P$ as:

$$\beta_p(L) = \frac{\text{card} \left( \text{POS}_p(L) \right)}{\text{card}(U)} = \frac{\text{card} \left( \bigcup_{i \in L} P_i(Y_i) \right)}{\text{card}(U)} \quad (8)$$

Definition 3.2 let $P, R \subseteq Q$ and $R \subseteq P$, if $\gamma_p(L) = \gamma_R(L)$ and $R$ is the minimum set of meeting demands of the equation, the $R$ is the reduction of $P$ denoting $\text{RED}(P)$. As may be inferred from this, the impend approximation measurement of classification $L$ about condition attribute set is not changed before or after the reduction.

2.4 Algorithmic steps

Let us suppose, there are $m$ attributes, $C_1, C_2, \ldots, C_m$ in the set $C$, the condition attribute set of the decision system; $D$ is the decision attribute set, the classification determined by $D$ is $\{Y_1, Y_2, \ldots, Y_k\}$; for every condition attribute $C_i$ we can calculate the $k+4$ parameters: $\alpha_{c_i}(Y_j), \gamma_{c_i}(L), \phi_{c_i}(L), \beta_{c_i}(L)$, in that, $i=1\sim m, j=1\sim k$. $\lambda_{c_i}$ and $\mu_{c_i}$ are the arithmetic mean and the geometric mean of these $k+4$ parameters. The condition attribute becomes more general and more logical to the importance of the decision classification by at the same time considering the absolute and relative classification of the condition attribute and the decision attribute in these $k+4$ parameters of every condition attribute $C_i$.

Definition 4.1 The definition of the importance of the attribute $C_i$ is:

$$Z_{c_i} = \alpha_1 \lambda_{c_i} + \alpha_2 \mu_{c_i} \quad (9)$$

The $a_1$ and $a_2$ is the importance parameters of the arithmetic mean and the geometric mean determined by user. When all of $K+2$ parameters are not zero, this attribute has an effect on all of subset. So, increasing the geometric mean $U_{c_i}$ is for expressing the importance of this effect.

The following is the detailed steps:

Step 1: Calculate the $\gamma_C(L)$ of the condition attribute set;

Step 2: Calculate the $Z = \{Z_{C_i}\}$ for one randomly selected condition attributes

Step 3: Initialize $C^0 = \phi$,

Step 4: $C^0 = C^0 + \{C_i | \forall i, C_i \text{can make } Z_{C_i} \text{ the maxmun}\}$

Step 5: Judge $\gamma_{C^0}(L) < \gamma_C(L)$, if the result is true, then turn to step 6, if the result is false, turn to step 4

Step 6: $C^0$ is one of least reduction

2.5. Example

In the table 1, the indexes of the telecommunication network quality evaluation could be the following 6 attributes, $C_1, \ldots, C_5, D$, meaning the network load capability, service area, information translation time delay, center address, congestion signal, quality of the telecommunication network.
Supposing the parameters $a_1$ and $a_2$ are 0.5, after the calculation, the importance parameters of every condition attribute are shown in the following table 2:

<table>
<thead>
<tr>
<th>Importance parameters</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z_{c_i}$</td>
<td>0.259</td>
<td>0.541</td>
<td>0.087</td>
<td>0.632</td>
<td>0.095</td>
</tr>
</tbody>
</table>

3. Results and Discussion

From the table 2, the least attribute set is $C^0$={ C1,C2,C4}. The attributes of the decision system is reduced to (C1,C2,C4) after eliminating the redundant attributes. According to the deduction theory of the classification rule, after eliminating the unnecessary attribute value in the decision table, the classification rule of the image decision is given as followed:

R1: (C1=Great)$\rightarrow$(D=good)  
R2: (C1=Middle)$\wedge$(C4=L2)$\rightarrow$(D=good)  
R3: (C1=Middle)$\wedge$(C2=common)$\rightarrow$(D=good)  
R4: (C1=Middle)$\wedge$(C2=large)$\wedge$(C4=L1)$\rightarrow$(D=good)$\vee$(D=bad)  
R5: (C1=Weak)$\rightarrow$(D=bad)  
R6: (C5=Not calling)$\rightarrow$(D=bad)  
R7: (C1=Middle)$\wedge$(C3=short)$\wedge$(C4=L1)$\rightarrow$(D=bad)

Result analysis

Finding the minimum knowledge reduction is an NP-hard issue; Its complexity is mainly caused by a combination of attributes in decision table. For the algorithm above, In the worst of conditions, the number of attributes to be considered for each times are: $m, m-1, \ldots, 1$ (the number of the condition attribute in decision table is represented by $m$ ), so the total is:

$$m + (m - 1) + \cdots + 1 = m(m + 1) / 2$$

(10)

So, if the influence of object number to computing time is ignored, thus, In the worst of conditions, the satisfying knowledge reduction can be find within a time complexity of $o(m^2)$ by this algorithm.
The innovation of this paper is to define the five parameters as follows: an approaching approximation measurement of a collection of Y on the property P, an approximation classified quality of decision-making division of L on the set of attributes P, a rough approaching approximation measurement of decision-making division of L on the set of attributes P, a class quality measure of decision-making division of L on the set of attributes P, and the relative classification of decision-making division of L on the set of attributes P. On this basis, the structure of the algorithm taking into account the absolute classification and relative classification of condition attributes and decision attributes, which makes the classification importance of condition attributes on decision, can be more comprehensive and reasonable.

4. Conclusion

In this paper, the data mining method based on the rough set is discussed, one kind of improved attributes reduction method based on the rough impedance approximation measurement is explained and applied in the telecommunication network quality evaluation. There are 2 advantages of this algorithm. One is that the attribute reduction from the algorithm of this paper is obtained by calculating the attribute importance of every condition attribute, besides, the dependent importance of the condition attribute to the decision attribute can be organized in order which is not drawn from the traditional attribute reduction methods. Another one is that this algorithm orderly put the attributes into the reduced attribute set according to the importance of different attributes from high to low until meeting the reducing requirement. So this algorithm is easier to realize. Especially when there are many condition attributes, the attribute reduction can be calculated faster. Comparatively the traditional reduction algorithm with the identification matrix needs more memory space because of the increased time-complexity of the algorithm in case of many condition attributes.

Acknowledgements

This study is supported by the Natural Science Foundation of Shaanxi Province (No.2012GQ8050)

References


