Application of Rough Set Theory in Intelligent Information Processing

Haoyue ZHU

School of Information Engineering, Xi'an University of Arts and Science, Xi’an 710065, China

Keywords: Intelligent information processing, Rough set, use.

Abstract. This paper will analyze the uncertainty, incompleteness and complexity in the current information processing technology. Using methods and principles of rough set theory, we analyze the functional use of this method in intelligent system and provide some new methods for the use of these technologies.

“Rough set” is a theory put forward by the Polish scholar Z. Pawlak [1], which solves the boundary problem existing in the "vague" theory of G. Frege [3]. This theory is different from the "fuzzy set" theory proposed by L.A. Zadeh [5]. There are two parts of the two questions it discusses: clear subsets and definable subsets. While, the clear subset includes the upper approximation set and the lower approximation set. These two elements mainly analyze the concept of "vagueness". It is to categorize those elements that are not sure, for example, the upper and lower approximation sets in the boundary domain are different. This facilitates the calculation of vague elements.

The Concept of Rough Set Theory

Rough set

The rough set is the R system with equivalence relation on the domain U, which constitutes the equivalence class. There is an own X⊆U, where the lower approximation $\underline{R}$（X）={x∈U：$[x]_R \subseteq$ X}

The upper approximation $\bar{R}$（X）={x∈U：$[x]_R \neq \phi$}

To say both inside and outside of X,$[x]_R$ means that there is an element x in the equivalence relation R. Then analyze according to the smallest knowledge "particle" established by R. Rough sets are brought into the rough function for analysis, which is: μ_X^R（x）= $\dfrac{card(X \subseteq [x]_R)}{card([x]_R)}$

In the above formula, μ_X^R（x） represents the non resolvable relation in R and the degree of binding of X in x. The equivalent elements in the same domain are the same. This is the rough membership function value, and 0≤μ_X^R（x）≤1。

The research on the logic of rough sets and the content of approximate reasoning include Rough logic, Rough set topology, Rough set algebra and so on. Among them, the Rough logic is analyzed from a topological perspective by T.Y.LIN,etc. He said rough sets and theoretical analysis are equivalent links, that is an approximating equivalence relation.

Lower Approximate and Upper Approximations in Incomplete Information Systems

S=（U，C，V，f）is an incomplete information table for X sets. And the attribute B$\subseteq$C means the upper approximation（X^{B}）and the lower approximation（X_{B}）. The definition is:

X_{B}={x$\in$U|$R^{-1}$$\subseteq$X}，

X^{B}=UR（x），x$\in$R

These show that if both x and asymmetric objects are in the X set, that is to say x belongs to the X class. Otherwise, if x is in an asymmetric similar object, it means that the object is an X class.

The Application of Rough Set in Intelligent Information Processing

Getting information automatically is to get what you want from the history. It is available on any system and also requires an analysis of preprocessing issues. Among them, the functions of collecting, sorting and sampling raw data, etc, need to be arranged in different ways. However, the original data will not be obtained directly. It is necessary to carry out certain caring and processing and add missing information in the original data. While,the original data range is real valued and needs to be discretized because the Rough theory can only select discrete values of objects. Related completion algorithm below:

Mean Completer Algorithm

This algorithm divides data into two types: non numerical attributes and numerical attributes. If the detected missing data is numeric, it can be filled according to the average value of the information attribute. If the check is non-numeric, it is necessary to give it a value on the instance (the frequency is very high), so that you can complete the filling of these data. Look at table 1.

Table 1 Incomplete data information table

U	Color String	Crade Float	Radius Float	Sold String	Year Integer
1	Red	1.0	3.14	No	1970
2	Green	1.5	2.71	Yes	1492
3	Red	2.0	10.65	Yes	1814
4	Red		0.98	No	
5	Blue	3.5	0.2	No	1776
6	Yellow	2.5		No	1865
7		6.0	4	Yes	1968

The Radius in Table 1 is an unknown property. And the padded way is（3.14+0.98+2.71+0.2+10.65+4925.6）/6=823.87. The color in 7 is also an unknown property, it is non numerical attributes. The "Red" is the number of occurrences more, it also needs to be filled. The complement of other algorithms is the same.

The Mean Completer is a straightforward algorithm that includes the Combinatorial Completer algorithm. The Combinatorial Completer also requires the data attribute when supplementing data, and then averaged, but it is not extracted from the information table, but from the same decision attribute value.

Combinatorial Completer Algorithm

Combinatorial Completer algorithm is another complement algorithm, which adds values according to the vacancy attribute. This algorithm is very simple, it needs to be based on a lot of data or attribute values and the amount of calculation is really great. However, there is a missing attribute in the Conditioned Combinatorial Completer algorith. But it analyzes all the values from the same instance, which reduces the cost of the Combinatorial Completer algorithm. There are 6 possible values for the Year attribute{1970,1492，1814,1776,1865,1968}. And Crade has 6 values, such as {1.0, 2.0,1.5,2.5,3.5,6.0} and so on, which can be filled by using 6 x 6 in Combinatorial Completer algorithm.

Discretization Decision Table

This decision table is analyzed using the Rough set theory, which requires data values to be discrete data. That is to say, decision attributes and range are connected, and discrete processing is needed before processing.

In fact, the problem of discretization is to classify the endpoint and the space of attribute conditions. We set a value n (it is the number of attributes of the condition) countless regions in space and the decision value of the region's objects is the same. If you select a m-1 on the endpoint, the attribute of this number grows with the breakpoint attribute. This is the combination of attribute values and breakpoints, which reduce the number of values according to the merged attributes, thus reducing the complexity, so that the corresponding values can be obtained.

Naive Scaler Algorithm

This algorithm analyzes the attribute $a\in C$, and analyzes according to a (x), arranges $x\in U$ from small to large, scans from top to bottom. Where x1g and x2 are adjacent instances. If a (x1) = a (x2), you need to scan again;. If d (x1) = d (x2) is the same decision, then scan again. Otherwise, you will get an endpoint c,c=(a（x1）+a（x2）)/2.

Naive Scaler algorithm does not use other parameters,it starts discrete processing according to the system information or database. It arranges attributes from large to small, and then judges and analyzes examples. If the decision and attribute values are not the same, the average value is chosen as the breakpoint at this time. This algorithm values the breakpoint based on the decision and conditional attributes, without considering the information relationship. And then finds different breakpoints according to different arrange ways of databases .

Semi Naive Scaler algorithm

It is not the same as the Naive Scaler algorithm. It handles the optional breakpoints of each attribute, and then analyzes whether these breakpoints can be analyzed.The specific analysis can be seen below:

The c attribute has a candidate breakpoint a. At this time x1 and x2 are two adjacent properties of c. And x1 <c, x2> c.

Let C1 be a set of a very large equivalence class decision in x1. When the two decision-making values appear the same frequency, we say | D1 |> 1.

If $D1\subseteq Dn$ or $Dn\subseteq D1$, do not choose this breakpoint. Otherwise choose this breakpoint.

It can be seen from the above analysis that the data obtained by the Semi Naive Scaler algorithm is smaller than that obtained by the Naive Scaler algorithm. It can remove the unnecessary breakpoints obtained by the Naive Scaler algorithm and reduce the breakpoints.

Rough Theory and Boolean Logic is Actually the Same Discretization Method

Algorithm Analysis:

Step 1: Gather all the attribute values.

Step 2: Use different matches to relate the properties between all the values.

Step 3:Analyze the two different decisions-making by using the above formulas.

Step 4:Choose the right way according to the format of these formulas.

Step 5: Normalized value processing.

Step 6:Choose a random value for discretization analysis.

Conclusion

According to the definition of the relationship between approximation and non resolution,rough set theory classifies people's knowledge that cannot be layered, and calculates them by software. This method has brought great help to intelligent information processing. However, the rough set theory needs further improvement and development.

Reference

[1] Z.Pawlak, Wang G, Wang X, et al. Application of Multi-Sensor Information Fusion Method Based on Rough Sets and Support Vector Machine[C]// Mechanical Engineering and Control Systems. Mechanical Engineering and Control Systems (MECS2015), 2016:350-353.

[2] Che Lin,Jiang Min.Application of rough set theory in freeway traffic incident intelligent matching system[J].Engineering Technology: Digest, 2016(12): 00249-00250.

[3] G.Frege, Kole D K. A Time Efficient Leaf Rust Disease Detection Technique of Wheat Leaf Images Using Pearson Correlation Coefficient and Rough Fuzzy C-Means[M]// Information Systems Design and Intelligent Applications. Springer India, 2016.

[4] Hao Jie,Shi Kebin, Wang Xianli, et al.Application of Cloud Model Based on Fuzzy C- mean Algorithm and Rough Set Theory in Rock Burst Rating[J].Geotechnical Mechanics,2016, 37(3):859-866.

[5] L.A.Zadeh, Kavitha V, Geetha T V. Influential Researcher Identification in Academic Network Using Rough Set Based Selection of Time-Weighted Academic and Social Network Features[J]. International Journal of Intelligent Information Technologies, 2017, 13(1):1-25.