International Journal of Database Theory and Application
Vol.9, No.11 (2016), pp.271-280
http://dx.doi.org/10.14257/ijdta.2016.9.11.24

A Study on Decomposition Method of Incomplete Information System Based on Data Mining Algorithm

Tu Pin

Hunan Vocational College for Nationalities, Yueyang City, Hunan Province, China, 414000
23196506@qq.com

Abstract

Recently network data domain knowledge updates quickly, but with the growth of the large amount of information, the stability of the information itself decreases dramatically. So, one of the key research directions is that how to dig out the valuable information from the unstable and chaotic huge information. The research on rules getting incomplete information is helpful for getting more useful information. When the incomplete information turns into complete, it will cause a certain degree of information distortion. For this problem, the paper proposes the decomposition method of incomplete information system. This method, without completion process of incomplete information, selects a template through a template function. The template function is based on the rough set theory, and when ensuring the template, it can extract subset from incomplete information through decreasing step by step. Incomplete information system need to use an intermediate variable based on rough set theory when it is broken down by simplified rule sets.

***Keywords**: data mining, decomposition method, incomplete information system, rough set*

1. Introduction

With the rapid development of information and, unstable data information is increased to a certain size. If people want to deal with the large, incomplete data, they have to adopt a kind of advanced technologies-data mining to identify and dig out the useful information [1]. Data mining is a method that it can dig out useful information from large information and incomplete information. It is a necessary process of actually building database. There are two main parts for the research of the incomplete information system. The one is that rules getting complete system. However this part has a disadvantage and loses a part of data information. The other part is that directly research incomplete information system [2]. Recently, in most cases, people will complete the incomplete information system for the processing of incomplete information system. But after completion, the knowledge gained in data mining will exist a certain degree of distortion. Meanwhile it is unbelievable that the system is completed when there are many null values [3]. Rough Set Theory is a kind of method to study incomplete information from data mining system. There is a advantage. The advantage is that after information processing, the original data is true [4]. In the process of data mining, if people want to define the interrelation, the interrelation can’t analyze, and the rough set theory is only used. Many information systems for the method of data mining are based on rough set simple theory, such as different information systems, fast developing data mining, range intelligent algorithm, information concept and so on [5]. In order to overcome the shortcoming of information distortion, the paper studies decomposition method based on incomplete information system. This method, without completion process of incomplete

ISSN: 2005-4270 IJDTA

information, selects a template through a template function. The template function is based on the rough set theory, and when ensuring the template, it can extract subset from incomplete information through decreasing step by step. Incomplete information system need to use an intermediate variable based on rough set theory when it is broken down by simplified rule sets [7].

2. The Data Mining Research Background

Computer develops rapidly especially in the hardware and software. With the popularity of database technology application, a large amount of data in the database is following. Facing the data ocean that expands rapidly, the problem, which how to find the useful method that in the huge database we can find the data of helping and serving people, has become the focus of the information technology researchers. Arose data mining is that in the huge database, digging out the association pattern that is helpful for scientific research and has the potential relationship. For the most of these data that is incomplete and uncertainty.So it is necessary to research the data mining in incomplete information system.

2.1 Incomplete Information System Research Situation

The research about incomplete information includes two mainly parts: rules getting and directly research incomplete information system. Below the paper will discuss in detail:

2.1.1. Rules Getting

Rules getting transforms the incomplete information system by rules getting method of complete information system. Once the incomplete information system turns into the complete information system, various theories and methods of the complete information systems can process and get rules. In the process of transformation, it needs to deal with the null value. The simplest way is to remove objects with the null value, but this method will cause waste data. The data can be filled, after transformation, it can be used the method of complete information system. However, there is a process of data analysis in the process of data filling. So, the relationship of other condition attributes and decision attributes is used to estimate the null value in the process. Generally, the methods are the Bayesian model, evidence theory, and the rough set theory and so on. Probability density and evidential functions are needed for the evidential functions and evidence theory. But data needs to be extra added and is hard to find. Therefore, the rough set theory becomes the focus of the researchers.

2.1.2. Directly Research Incomplete Information System

It researches the incomplete information system without changing the information system.

Because there will be the uncertain value, it is hard to find the needed equivalence relation in the object set. It means we can only consider the similarity among objects. Recently, the research of incomplete information system mainly adopts the tolerance relation, non-symmetric similarity relation in the rough set theory.

2.2. The Basic Concept of Rough Set Theory

In the early 1980 s Poland scientist Pawlak proposed the rough set theory. This theory is mainly used for data analysis, and provides a new mathematical method especially for the inaccurate incomplete data. The rough set theory only needs to delete redundant data according to observation data and compare the degree of incomplete data. It doesn’t need any priori information. Rough set has been successfully used in machine learning, decision analysis, and process control and so on.

The definition of rough set:

Rough set theory can define an information system for four-place combination:

$$P=（U,N,T,f） \quad (1)$$

P=（U,N,T,f）is nonempty limited set,H is the equivalence relation of U,Y⊆U. So, we can get:

The upper approximation of Y:

$$H^{-}(Y)=\{y\,|\,(y\in U)\wedge([y]_R\cap Y\neq\Phi)\} \quad (2)$$

The lower approximation of Y:

$$H_{-}(Y)=\{y\,|\,(y\in U)\wedge([y]_R\subseteq Y)\} \quad (3)$$

The boundary region of Y:

$$BI_R(Y)=R^{-}(Y)-R_{-}(Y) \quad (4)$$

If $BI_R\neq\Phi$, Y is the concept of rough set. The values of lower approximation contain all elements exactly classified Y.The values of upper approximation contain all elements that maybe belong to Y.The boundary region is a region set that remove the elements that maybe belong to set Y and extra set Y is in the knowledge set R and U.

3. Decomposition Algorithm

The decomposition algorithm of incomplete information system is a method that deals with huge complex data even missing data. Compared with other methods, there are two advantages: the one is that it reserves the truth of factual data, the other one is that it doesn't have additional information because of completing the incomplete information system. The decomposition algorithm is widely used in many fields. The decomposition algorithm of incomplete information system is mainly used in the database decomposition algorithm. This decomposition algorithm can minimize the influence of incomplete information. There are two keys about rough set decomposition algorithm of incomplete information system: the one is choosing the templates what meet the requirements. The other one is disintegrating incomplete information system. Decomposition algorithm is very simple; it can turn the complex and incomplete information into simple and easy handle information. Two important steps are building a template evaluation function and intermediate variables. Firstly, proposing a template evaluation function based on rough set theory. This template function contains few properties, but it can still maintain a high degree of classification ability. Secondly, using the concept of rough set to analyze the middle concept, the incomplete information system is disintegrated and classified by the intermediate variable.

The paper will explain an important concept before introducing decomposition algorithm, in order to better understand the process of decomposition algorithm. $S=(U,A,V_a,a)$ is the information system, a is a mapping of U to V_a .If there is $x\in U$ and there isn't the value of $a(x)$,the S is incomplete information system.

3.1. Proposing the Evaluation Function

The most important step for incomplete information system decomposition method is building an optimal template construction of evaluation function. For a factual problem(a given system),we should consider the following aspects:(1)the optimal decomposition algorithm of the given problem(based on rough set which is the most widely),(2)how to build the optimal template construction of evaluation function after decomposition algorithm,(3)can all the practical problems find optimal template function?

The most effective data deformation method is database decomposition algorithm. Through database decomposition method can reduce the influence of missing value for incomplete information system data mining on a certain extent. We can identify a subset $U_{C_1} \subseteq U$ from the original incomplete information system based on evaluation function. It means that the subset can keep the primitiveness of data. So, it can completely used for data mining in incomplete information system. Obviously, the selection of template function is very important; it is related to the accuracy of the whole algorithm. So building the optimal template evaluation function is vital for decomposition algorithm and this is the focus in the practical application. The identification of decision table subset and the quality of the classification are based on template. So the researches need to define a new evaluation function. Based on this new function, the researchers can choose a smaller subset from attribute subset quickly and preferentially. The function is following:

$$\vartheta(t) = {}^{v(C_t\{d\})}\!\Big/_{card(C_t)} = \frac{card(POS_{ct}(d))}{card(C_t)\cdot card(U_{c1})} \qquad (5)$$

$POSc_t(d)$ is a collection of elements from correctly classified $U/\{d\}$ base on C_t. $card(\cdot)$ is the number of existing elements in the collection. $\upsilon(C_t,\{d\}$ is classified ability of subset based on C_t. $card(C_t)$ is the number of attributes in the subset.

By analyzing the construction of evaluation function, it finds that it's difficult to find the optimal template. There is the optimal template function, but in fact, there isn't. So in actual application, people find qualified template instead of optimal template function to settle for second best. People always use an algorithm (heuristic) and give the threshold ϑ_0.Wrapper method is a way to search, starting from zero attribute and increasing one by one to make $\vartheta(t)$ larger until $\vartheta(t) \geq \vartheta_0$.Then immediately put an end to the search. So people can get the required template. It has been found that the template evaluation function through this method can user fewer attribute variables to classify the information, and greatly reduces the difficulty of the operation.

3.2. Building the Intermediate Variable

After identifying the subset $U_{C_1} \subseteq U$ without the missing values through template evaluation function concept, the attribute of attribute subset C_t is more than one. If the database decomposition is a multi-attribute simple conjunction, this will cause many complicated database subset. Sometimes it even causes the data fitting problems. Therefore, a new intermediate variable IB based on rough set theory is needed, to decompose the database.

$$\mathrm{U_{C_t}} / \mathrm{IND(C_t)} = \{\alpha_1, \alpha_2, \cdots \alpha_n\} \qquad (6)$$

$$\mathrm{U_{C_t}} / \mathrm{IND(d)} = \{\beta_1, \beta_2, \cdots \beta_n\} \qquad (7)$$

A intermediate variable IB is the new equivalence relation of U.

$$\left\{\begin{aligned} &\mathrm{U/\,IND(IB)} = \{\mathrm{IB_1}, \mathrm{IB_2}, \cdots, IB_m, IB_{m+1}\} && (8)\\ &\qquad \mathrm{IB_{m+1}} = \mathrm{W} \cup (\mathrm{U} - \mathrm{U_{C_1}}) && \\ &\mathrm{IB_i} = \cup\{\alpha_j \in \mathrm{U_{C_t}/IND(C_t)} : \alpha_j = \beta_i\} && (9) \end{aligned}\right.$$

In the $\mathrm{W} = \cup\{\alpha_j \in \mathrm{U_{C_t}/IND(C_t)} : \alpha_j \not\subset \beta, \mathrm{i}\}, \mathrm{i} = 1,2,\cdots,\mathrm{m}$, we can find C_i can correctly decompose subset IB to $IB_i(1 \leq i \leq m)$.There are two parts in the IB_{m+1} : the one is excluded object. This kind of object is ruled out $U_{C_t} \subseteq U$ because containing the null values. The other object is that it can't be correctly classified with C_t .

Introducing the intermediate variable can ensure the selected attribute subset contains the necessary attributes, and also can avoid that the decomposed attribute subsets are too many or too complex. The information set by dealing with intermediate variable is complete but not repeat, and it is more convenient for actual operation.

3.3. The Process of Database Decomposition

The process of database decomposition is shown in picture1.For actually received incomplete information system, first of all, the corresponding database can be found, and several attribute subsets of containing attribute can be got through Database decomposition method. Then the eligible template selected by template evaluation function is not optimal. In order to avoid the disadvantage that the attribute subset is too many or too complicated, people always introduce an intermediate variable to take the equivalent transformation with equivalent transformation. At last, people should observe the number of sub database elements.These database elements is transformed with intermediate variable. If the number is zero, the result meet requirement. If the number is not zero, the databases will instead the original incomplete information system to repeat the above process.

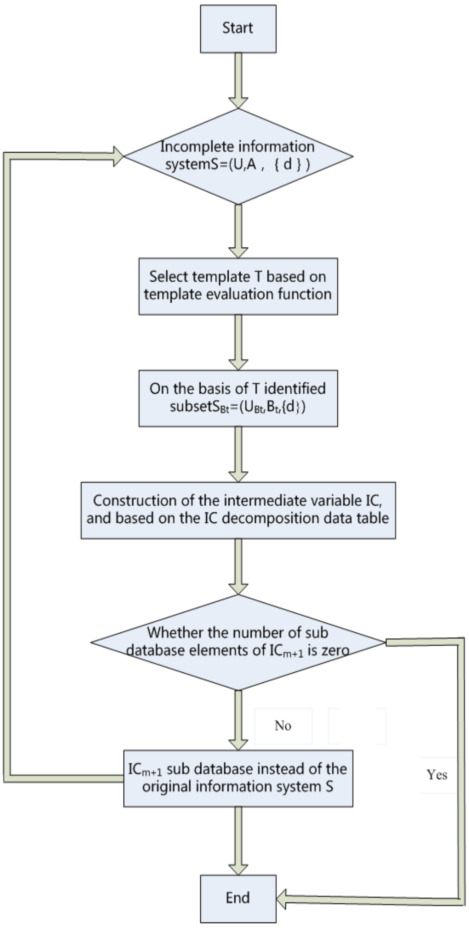

Start
Incomplete information systemS=(U,A，{ d })
Select template T based on template evaluation function
On the basis of T identified subsetS$_{Bt}$=(U$_{Bt}$,B$_{t}$,{d})
Construction of the intermediate variable IC, and based on the IC decomposition data table
Whether the number of sub database elements of IC$_{m+1}$ is zero
No
Yes
IC$_{m+1}$ sub database instead of the original information system S
End

Figure 1. Algorithm Proces

4. The Example Verification of Decomposition Algorithm

Here, the paper takes the vibration fault diagnosis problem that it is common in the gas turbo generator. The effectiveness of rough set decomposition algorithm of incomplete information system can be verified with this example. Firstly, reasonably classify the data, and get the decision rules set. $U_{C_t} \subseteq U$ Fault diagnosis table (incomplete information)is shown in Table 1.Clearly,there is missing in the actual data information.* is the missing part. Obviously, the incomplete information system is composed by the common vibration fault diagnosis problem of the gas turbo generator and related data. How to use decomposition algorithm to dig out data based on incomplete information system will be explained.

Table 1. Decision Making for Incomplete Fault Diagnosis

U	G	H	J	K	L	P
1	0.053	0.783	0.222	*	0.014	1
2	0.231	0.977	0.313	0.055	*	1
3	0.163	*	0.283	0.024	0.016	1
4	0.026	0.064	0.981	*	0.056	11
5	0.043	0.023	*	0.315	0.065	11
6	0.011	0.052	0.874	0.182	*	11
7	0.032	0.035	0.388	0.532	0.230	111
8	*	0.026	*	0.459	0.105	111
9	0.015	*	0.428	0.497	0.175	111

The attributes G, H, J, K, L are gas turbine generator unit vibration signal frequency domain characteristics of the spectrum $< 0.4f, 0.4f \sim 0.5f\ , 1f\ , 2f\ , \geq 3f$, (f represents frequency), the amplitude component energy of five spectrum. Decision attribute P is the fault category of gas turbine generator. Its value is 1, 11,111, respectively represent three common faults of gas turbine generator. These faults are the oil film surface oscillation, balance disorder and misalignment.

(1)data preparation process

The processing of decision table is based on the rough set theory. This is because the attribute values are expressed by the discrete values. so the Table 1 is deal by discrimination, then the discrimination Table 2 is got.

Table 2. Discrete Fault Diagnosis Decision Making

U	G	H	J	K	L	P
1	1	1	0	*	0	1
2	1	1	0	0	*	1
3	1	*	0	0	0	1
4	0	0	1	*	0	11
5	0	0	*	0	0	11
6	0	0	1	0	*	11
7	0	0	0	1	1	111

(2)the template selection

The template selection is a template evaluation function according to formula 1.The subset of U_{C_t} decided by template of $C_{tGJ} = \{G, J\}$ has higher classification ability.

$\vartheta(t_{GJ}) = 10/(10 \times 2) = 0.5$, so using the template of t_{GJ} for decomposing the incomplete information system, the subset $(U_{B_{tGJ}}, B_{tGJ}, \{P\})$ is got.

(3)building intermediate variable

On the $(U_{B_{tGJ}}, B_{tGJ}, \{P\})$, using formula 5, intermediate variable IB_1 can be built. So the first layer of diagnosis rules set shown in the following table.

Table 3. First Tier Diagnostic Rules Set

IB_1 Value	Rule
IB_1=1	G=1,J=0→P=1
IB_1=11	G=0,J=1→P=11
IB_1=111	G=0,J=0→P=111

The object without processing will be classified into $IB_1 = 4$ (shown in the following table), then finish the decomposition and get the rules.

Table 4. IB1=4 Sub Data Table

No.	G	H	J	K	L	P
4	1	1	*	0	0	1
5	*	1	0	0	0	1
7	0	0	*	0	0	11
9	*	0	1	0	0	11

(4)classification of the data set

Through analyzing the Table 4, containing the missing value objects of the original information has been classified correctly after using the first layer rules. In other words, the influence of the missing values has been overcome when digging out the data.

Then the same process, selecting the template in the sub data $B_{tHK} = \{H, K\}$, the subset $(U_{B_{tGJ}}, B_{tGJ}, \{P\})$ is met the requirement. Based on intermediate variable IB_2 built by subset, the rules set is about the diagnosis of the second layer. Shown in the following table:

Table 5. Diagnosis Rules Set

IC_2 Value	Rule
IB_2=1	H=1,K=0→P=1
IB_2=11	H=0,K=0→P=11
IB_2=111	H=0,K=1→P=111

By analyzing the actual vibration fault diagnosis problem of the gas turbo generator, it finds that there are some properties and characteristics based on the incomplete information system decomposition algorithm as following:

(1)Because the process doesn't need to complete the incomplete information system, the data used by decomposition algorithm can keep the data authenticity and data primitiveness. The study results conform to the actual; there is the high degree of accuracy. To a certain extent, it can reduce the waste of information data, there isn't additional redundant information, and it also avoids losing authenticity.

(2)There is an advantage in this method. The advantage is simulating artificial intelligence. The way of dealing is similar to human in a strange environment. The way of dealing with strange information is that the strange information or something is classified according to the level and then identified one by one. The thinking of this decomposition algorithm is similar to the human brain thinking. It will classify, discuss one by one, and it is easier to be understood and accepted. It will get the more accurate results when handling the actual problems.

(3)After confirming the classification model about decomposition algorithm, for the new object, the process is from top step by step matching. And the new object will be taken into the subset table at the same time. This way has high extensibility and dynamic characteristics.

(4)Through the actual example, it finds that decomposition algorithm of incomplete system has higher practicability. The process is clear, the operation is simple and the results conform to the actual situation better.

5. Conclusion

In the field of intelligent information processing, data mining in the incomplete information system is the key and difficult points for a large number of researchers. The incomplete data dealt with complete method can't avoid a certain degree of distortion. If the number of objects containing null values is too much, complete system is not feasible and there are some disadvantages. If the actual data is too much or is missing, the data mining is a common method. In many data mining processing method, there are many methods for the incomplete information system. These methods are roughly divided into two aspects: the one is that complete the incomplete information system, and then deal with the complete system; the other one is that directly deal with the incomplete information system. According to the above situation, the paper proposes the decomposition algorithm of incomplete information system based on data mining. Based on the rough set theory, the decomposition algorithm can dig data for incomplete information system. In order to verify the effectiveness of this method, the paper explains the vibration fault diagnosis data about the gas turbo generator. This method removes the process of completing incomplete information system. Because the process is similar to human intelligence, this method can keep good extensibility and dynamic characteristics. This method in the field of data processing in a certain extent promotes the research progress. It can directly dig data for incomplete information system and keeps the original data.

References

[1] L. Hongli, W. Jue and Y. Yiyu, "User-oriented feature selection for machine learning", The Computer Journal, vol. 50, no. 4, **(2007)**, pp. 421-434.

[2] W. Ling, L. Hongru and Z. Wenxiu, "Knowledge reduction based on the equivalence relations defined on attribute set and its power set", Information Sciences, vol. 177, no. 15, **(2007)**, pp. 3178-3185.

[3] Y. Y. Yao and Z. Yan, "Attribution reduction in decision-theoretic rough set models", Information Sciences, vol. 178, no. 17, **(2008)**, pp. 3356-3373.

[4] Y. Y. Yao, "Concept lattices in rough set theory", In: Proceedings of 2004 Annual Meeting of North American Fuzzy Information Processing Society, Canada, **(2004)**, pp. 796-801.

[5] Z. Wenxiu, W. Weizhi and L. Jiye, "Theory and Method of Rough Set", Beijing: Science Press, **(2001)**.

[6] K. S. Qu, Y. H. Zhai and J. Y. Liang, "Study of decision implications based on formal concept analysis", International Journal of General Systems, vol. 36, no. 2, **(2007)**, pp. 147-156.

[7] K. S. Qu and Y. H. Zhai, "Generating complete of implications for formal contexts", Knowledge-based Systems, no. 21, **(2008)**, pp. 429-433.

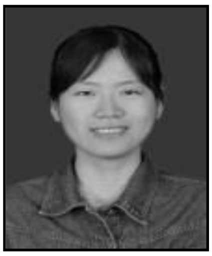
[8] Y. Leung and D. Y. Li, "Maximal consistent block technique for rule acquisition in incomplete information systems", Information Sciences, vol. 153, **(2003)**, pp. 85-106.

[9] X. L. Hunag, "A pseudo-nearest-neighbor approach for missing data recovery on Guassian random data sets", Pattern Recognition Letters, vol. 23, **(2002)**, pp. 1613-1622.

[10] J. R. Quinlan, "C4: programs for MaehineLearning", Morgan Kaufman, SanMateo, **(1993)**.

[11] A. Skowron and C. Ranszer, "The diseernibility matrices and functions in information systems", InR. Stomwinski, editor, Intelligent Derision Support. Hand book of applications and Advances in Rough Sets Theory, Dordreeht, Kluwer, **(1992)**, pp. 331-362.

Author



Tu Pin, is born in June, 1981 in Yueyang, Hu'nan Province. She is a master. Her major is Computer Science and Application major. She works at Hunan Vocational College for Nationalities in Yueyang City; Hu'nan Province now.She hosts a project named Supported by Natural Science Foundation of Hu Nan Province of China: Application and practice of Association rule mining in Personalized Service Based on Sakai network platform.15C0646.