A Novel Algorithm for Filling Incomplete Data Based on Attribute Reduction

Yang Yang
School of Computer & Information Science,
Southwest University[1]
Chongqing Engineering Research Center for Instrument and Control Equipment[2]
Chongqing, China
yycia@swu.edu.cn

Shan Xiong Chen
School of Computer & Information Science,
Southwest University
Chongqing, China
csxml@163.com

Feng Liu
Chongqing Engineering Research Center for Instrument and Control Equipment
Chongqing, China
liuf@swu.edu.cn

Xiao Wen
School of Software Technology,
Dalian University of Technology
Dalian, China
swuwenxiao@gmail.com

***Abstract*—Traditional data filling algorithms take the unified method to fill the incomplete data, leading to the low efficiency for filling a large number of incomplete data. This paper proposes a novel algorithm for filling incomplete data based on attribute reduction. The presented algorithm first uses the attribute reduction to distinguish the important attribute from unimportant attribute, and then fills the two kinds of data attributes using different filling technologies. Experiments show that this algorithm can effectively and quickly fill the large-scale incomplete data.**

Keywords-attribute reduction; incomplete data; data filling

I. INTRODUCTION

With the development of the computer technology and the network technology, the number of data is increasingly large. Recent years, with the rapid development of the Internet of Things, more and more data collected by all kinds of sensors are incomplete, in other words, the problem caused by incompleteness of data is more and more prominent. Some attributes of data lose in the course of data collection due to the abnormal work of the terminals. The incompleteness of data brings tremendous difficulties to data fusion and data mining, impeding the development of computer application, especially the Internet of Things. Thus, filling the incomplete data effectively is an important topic.

Commonly used methods for filling incomplete data include the method based on decision tree, the method based on the Mahalanobis distance and the method based on the Bayesian networks and so on[1]. These traditional methods use the same technology to fill all the lost attributes, ignoring the different importance of the different data attributes, leading the low efficiency in the process of filling the incomplete data.

For the problem above, the paper proposes a novel algorithm for filling the incomplete data based on attribute reduction. The presented algorithm first uses the attribute reduction to distinguish the important attribute from unimportant attribute, and then fills the two kinds of data attributes using different filling technologies.

In the all of methods for attribute reduction, including the method based on heuristic[2] and the method based on decision table[3-4] and so on, the searching algorithm for attribute reduction based on power graph[5] is the most widely used method due to its high efficiency and convenience. However, this method is only used for attribute reduction of complete information system. So, the paper first gives the definition of the partition of the incomplete information system, which can convert the incomplete information system to the complete information. Then the method based on power graph is used for attribute reduction.

After getting attribute reduction, the paper improves the method for filling incomplete data based on similarity to fill the important attribute of the incomplete data. Similarity-based approach is used to define the compatibility relations and classification. However, this method will produce a lot of zero value of the object similarity, causing the serious imprecision of filling the incomplete data. This paper defines the similarity as the summation of the probability, deleting the interference of some abnormal zero value, improve the precision of filling the important attributes.

Experiments show the presented method can effectively and quickly fill the large-scale incomplete data.

II. RELEVANT KNOWLEDGE

A. Rough Set and Granular Computing

Definition 1. $IS = (U, A, V, f)$ is a information system, where U is the domain, namely the set of researched objects, A is the set of attributes, including condition attributes expressed by C and decision attributes expressed by D, V is the set of values of objects in U meeting the attributes A and f is the information function. When $x \in U$ and $a \in C$, $f(x,a) = *$ shows that the functional value of $f(x,a)$ is

unknown, which is called incomplete information system containing this kind of function, represented by IIS.

Definition 2. For an incomplete information system $IIS=(U,A,V,f)$ and $\forall B\subseteq C$, $DIV(B)$ defines a binary indistinguishable relation:

$DIV(B)=\{(x,y)\in U\times U \mid \forall a\in B, f(x,a)=f(y,a)$ $and(f(x,a)=*)\neq(f(y,a)=*)and(f(x,a)=*)\neq(f(y,a)=V_a^y)\}$, where $\times$ represents Cartesian product.

Definition 3. For a incomplete information system $IIS=(U,A,V,f)$, $U/DIV(C)=\{U_1,U_2,...,U_n\}$, where C is the set of condition attributes, its granularity is defined as: $GD(C)=\sum_{i=1}^{n}\frac{|U_i|^2}{|U|^2}$.

B. Power Graph

Definition 4. $Power(C)$ is the power set of C, the set of the condition attributes, and for the given directed graph G, the elements of $Power(C)$ are the vertices of G and the edges of G meet the next condition: for $\forall X,Y,Z\in Power(C)$,if $|X|-1=|Y|=|Z|+1$ and $(X\bigcap Z)\subset Y\subset(X\bigcup Z)$, there must be the directed edges from X to Y and from Y to Z. We call the directed graph G as the power graph of C.

C. Algorithm for Filling Data Based on Similarity

Definition 5. For a given incomplete information system $IIS=(U,A,V,f)$, where $x\in U$, $y\in U$, $A=C\bigcup D$, $a\in C$ and C is the set of condition attributes and D is the set of decision attributes. V_a is the value range of a. The probability, $P_a(x,y)$, that x and y get the same value is defined as follows.

(1) $1/|V_a|$, $f(x,a)$ and $f(y,a)$ are not simultaneously $*$;

(2) $1/|V_a|^2$, $f(x,a)$ and $f(y,a)$ are simultaneously $*$;

(3) 1, $f(x,a)=f(y,a)$;

(4) 0, others;

According to the definitions above, the formula for calculate the similarity is as follows.

$Similarity(x,y)=\sum_{i=1}^{n}P_a(x,y)$, where n represents the number of the attributes.

III. The Algorithm for Filling Incomplete Data Based on Attribute Reduction

The paper proposes a novel algorithm for filling the incomplete data based on attribute reduction. The presented method first uses the improved searching algorithm for attribute reduction based on power graph to abstract the important attributes of the data, and next uses the improved method for filling the incomplete data based on the similarity to fill the important these attributes. At last, the method for filling the incomplete data based on probability is used to fill the non-important attributes.

A. The Improved Searching Algorithm for Attribute Reduction Based on Power Graph

Input: incomplete information system

Output: the attribute set including *K* important attributes

Steps:

1 Create a search graph *G*, put the condition attribute set *C* into the extended node table *Search*, calculate the knowledge granularity of the starting node, *g = GD (C)*, so that *minGD = g*.

2 Create a extended node table of *Searched*, with the empty initial values.

3 Let *P=C*.

4 Loop: if the table, *Search*, is empty, exit the loop.

5 Select the first node from the table, *Search*, and remove it to the table, *Searched*. This node is called *n*.

6 Extend the node *n* according to the power graph and produce the temporary set of successor nodes and then add these successor nodes into the power graph *G*.

7 Let $P\in TEMP$, if $|P|<K$, break.

8 If a node $T\in TEMP$ and $T\notin G$, set a pointer to the node *n* and calculate the granularity of the attributes of the nodes in the *TEMP*. If there is a node *p*, $P\in TEMP$, *GD(P)=minGD(TEMP)*, put the node *S* with *GD(P)>minGD(TEMP)* into the table *Searched*, and put other nodes into the table *Search*.

9 Reorder the table *Search*.

10 Go to loop.

11 Compare the granularity of the attributes set of the nodes with *K* attributers in the table *Search*. Select *K* attributes with lower granularity.

B. Algorithm for Filling The Important Attributes Based on Similarity

1 Use decision attribute *D* to divide *U*, $U/D=\{X_1,X_2,...,X_n\}$

2 Put U/D in the table *Open*.

3 Loop: if the table, *Open*, is empty, exit the loop.

4 Select the first node from the table *Open*, namely X_i.

5 If one attribute of the object x in X_i belongs to *RED(K)*, use the method based on similarity to fill this incomplete attribute.

6 If one attribute of the object x in X_i does not belong to *RED(K)*, use the method based on probability to fill this incomplete attribute.

7 After filling the objects in X_i , remove X_i from the table *Open* to the table *Closed.*

8 Go to loop.

IV. Experiment Simulation and Analysis

The experimental data is the real-time monitoring data collected from digital home lab[6].

Assumption that the information system *IIS=(U ,A, V, f)*, $A = C \bigcup D$, *U={Rm1, Rm2,…, Rm10}, C={temp, humi, lumi, power, location}*, the condition attributes is for short as: *C={t,h,l,p,L}*, and *D={sensitivity}*. The detailed data is shown as the table 1.

Table 1.Experimental data.

U	C					D
	temp	humi	lumi	power	location	sensitivity
Rm1	28	41	200	31.4	S	3
Rm2	28	41	200	87.1	S	3
Rm3	29	42	*	43.5	S	2
Rm4	27.5	*	170	29.4	N	2
Rm5	29	42	*	*	E	1
Rm6	29	41	190	31.4	S	3
Rm7	28	42	170	29.4	N	2
Rm8	29	41	190	43.5	S	2
Rm9	28	42	180	32.7	E	1
Rm10	29	41	180	32.7	E	1

We assume that there three important attributes in our experiment. First the presented algorithm executes the method based on power graph to find these important attributes and the searching process is shown as the figure 1.

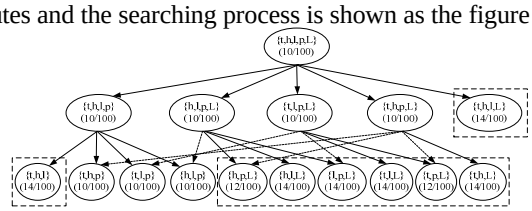

Fig.1 Important attributes searching

In the figure 1, the attributes are included in the brace and the granularity of each attribute is included in the parenthesis. The nodes in the red edging are not searched longer in their layer. From the figure 1, *{t,h,p}*、*{t,l,p}*、*{h,l,p}* are the attribute sets that Meets the requirements. We arbitrarily selected a set, such as *{t,h,p}*, as the important attribute set, namely *RED（3）={t,h,p}.*

Next, the algorithm uses decision to divide *U*: *U/D={{Rm1,Rm2,Rm6},{Rm3,Rm4,Rm7,Rm8},{Rm5,Rm9,Rm10}}* and then processes each division. The incomplete attributes will be filled according to the presented algorithm.

According to the proposed algorithm, we can get the values of the incomplete attributes as follows.

The value of the *lumi* of *Rm3* is 170.

The value of the *power* of *Rm9* is 32.7.

The value of the *lumi* of *Rm5* is 180.

The experimental analysis shows that the presented algorithm can quickly fill missing values of the incomplete attributes effectively.

V. Conclusion

With rapid development of the Internet of Things, more and more terminals are joined into the IoT, which work in no manual monitoring state, leading to destruction frequently. So a large number of collected data are incomplete, which brings difficult to data mining and data fusion and a serious impediment to the application of the Internet of Things data. For this problem, this paper proposes the paper proposes a novel algorithm for filling the incomplete data based on attribute reduction. The presented algorithm first uses the attribute reduction to distinguish the important attribute from unimportant attribute, and then fills the two kinds of data attributes using different filling technologies. The presented algorithm uses the improved method based on similarity to fill the important attributes, which can get high accurate missing values, and uses the method based on probability to fill the non-important attributes, which can get missing values in real time.

Acknowledgment

Our work is supported by (1) “the Fundamental Research Funds for the Central Universities” (XDJK2011C075). (2). the National High Technology Research and Development Program("863"Program) of China(2012AA041101) (3). “the Fundamental Research Funds for the Central Universities” (XDJK2011C075 2012-1-1)

We also would like to thank Chongqing Engineering Research Center for Instrument and Control Equipment provide equipment for our work.

References

[1] LI Hong, EMMMANUEL Amani, LI Ping, et al. Imputation algorithm of missing values based on EM and Bayesian network[J].Computer Engineering and Application,2012,46(5):123-125.G.

[2] HUANG Zhi-guo, Wang Duan. Study on data reduction algorithm based on rough set[J].

[3] CHEN feng-juan. Methods for calculating core attributes of inconsistence decision table[J].Computer Engineering and Design,2012,33(3),1187-1191.(3),1187-1191.

[4] GE Hao, YANG Chuan-jian, LI Long-shu. Efficient algorithm for computing core attributes[J].Computer Engineering and Application,2012,46(26):138-141.

[5] CHEN Yu-ming, MIAO Duo-qian. Searching algorithm for attributes reduction based on power graph[J].Chinese Journal of Computers,2009,32(8):1486-1492.

[6] Yang Liu, Zhikui Chen, Hao-zhe Wang, Xiao-ning Lv. An Architecture of Data Processing using Deluge Computing in Internet of Things[C].Dalian, China: IEEE International Conferences on Internet of Things, and Cyber, Physical and Social Computing.2011: 692-697.