A Data Cleaning Method Based on Association Rules

Weijie Wei[1] **Mingwei Zhang**[1] **Bin Zhang**[1] **Xiaochun Tang**[2]

[1]College of Information Science and Engineering, Northeastern University, Shenyang 110004, P. R. China
[2]Department of telecommunications, NEUSoft Group Ltd, Shenyang 110004, P. R. China

Abstract

The quality of the data affects the usability of the data mining's results. Making a data preparation before the mining can improve the quality. If the data are collected from the multi-data source, data preparation becomes very difficult. In this paper, a data-cleaning method based on the association rules is proposed. The new method adjusts the basic business rules provided by the experts with association rules mined from multi-data sources, and generates the advanced business rules for every data source. Using this method, time is saved and the accuracy of the data cleaning is improved.

Keywords: Data mining, Data cleaning, Business rules, Association rules

1. Introduction

Using data mining, one can discover the useful knowledge hiding in the plenty of data. More and more attention is paid to it. In order to get the worth knowledge, many researches have been made in the relate aspects [1], and the problem of data quality is one of the key ones.

For the reasons like fault input and missing the restrict conditions, there are dirty data in the collected data [2]. And the dirty data lower the quality of the data. For garbage in and garbage out, the quality of data affects the quality of the mining's results. Data cleaning is the step to deal with the fault data. It can improve the quality of the data.

Generally, there are two ways to clean the data. (1) Cleaning the data by the experts manually: Checking the datum one by one, all fault data can be detected and adjusted by the experts [3]. (2) Cleaning the data through rules or data models: the method described in document [4] cleans the data through the rules that extracted from the domanial knowledge by the experts. The method described in document [5] cleans the data through the rules that extracted from the sample of the data by the experts.

But, when the data collected from multi-data sources in large number, data cleaning methods above become unusable. If clean the data by the experts manually, the time and manpower spent on it will be beyond tolerance. Cleaning the data using data rules or data models seems to be reasonable and feasible. But, considering the difference among the multi-data sources, making the uniform rules or models can not satisfy the requirements. For example, in a data mining of Traditional Chinese Medicine Syndrome Differentiation (TCMSD) [6], the mining's data are collected from different cities: Shenyang, Dalian, Nanking, Shandong, and Guangdong. The people's physiques are different for the different existent surroundings in different cities. And the difference makes the concept of fever have different definitions. The air temperature of Shenyang is lower than it of Guangdong. Thus, the normal temperature of people in Shenyang is lower than it in Guangdong. In Shenyang, the average temperature is 36.5℃, one person with the temperature of 36.6℃ is regarded as a sicker with a low fever. And in Guangdong, the average temperature is 36.8℃, and then the person with the temperature of 36.6℃ is regarded as a healthy man. Assuming fever is the only attribute for diagnosis, it means that the person with no fever is a health one; otherwise the person is a sick one. Then the person with the temperature of 36.6℃ is diagnosed as illness in Shenyang and is diagnosed as health in Guangdong. Using uniform standard to clean the data from the different data sources all, some correct data will be cleaned or some incorrect data will not be cleaned. It affects the data quality severely.

The best way of solving the problem above is to make the rules for every data source. But because of the complexity in the data, it is impossible for the experts to make the all rules. Assuming there is a data set collected from N data sources, and every data source has M attributes, and every attribute associates with G (M) attributes in average; K is the number of difference between the data sources, and every difference affects H (K) attributes in average. (Here, N and M are natural number, and N is the number of the data sources, M is the number of the attributes in one data source, G (M) is a line shape function of M). Then the number of rules that the experts need to make is K*H (K)*G (M)*N, and when M, N and K increase, the value of the expression increases rapidly. So making the rules or models all by the experts is an impossible thing.

The method proposed in document [7] offers a new thought to solve this problem. The method detects, quantifies and corrects data though the association rules mined from the data themselves. In the document, an approach named DQM with association rule (DQMWAR) is also proposed. It improves the data quality only by the data themselves and depends on the experts less. Thus, the accuracy of the method depends on the number of the data in the database. When the scale of the database is not large enough, the accuracy is poor.

This paper presents a new data cleaning method based on association rules. The method cleans the data collected from multi data sources in a high accuracy. The rest of the paper is organized as follows. Basic idea and definitions are presented in section 2. Section 3 introduces the algorithm of integrating the association rules into advanced business rules. In section 4, the data cleaning algorithm based on association rules is proposed. Results from experiment are given in section 5 and section 6 deals with conclusions.

2. Basic idea and relate definitions

2.1. Basic idea

Basing on the basic idea of DMQ mentioned in document [7], a data cleaning method based on association rules is proposed. The method based on the association rules mined from the sample data, makes explore in integrating the association rules into the advanced business rules. The association rules are the relationships among the items mined from the data; the advanced business rules are the rules that the data must obey. Integrating the association rules, we can get the mining business rules. Merging the mining business rules with the basic business rules, we can get the advanced business rules. Using advanced business rules in the data cleaning process, we can clean the data effectively.

The basic framework of the method is showed in the Fig.1 on the right:

Using this method to do the cleaning, plenty of manpower would be saved for the method depends on the experts less. The only thing the experts should do is to make the basic business rules. The mining business rules are integrated from the association rules that mined from the data. And the accuracy [4] of the cleaning method is improved for the rules are based on the data themselves.

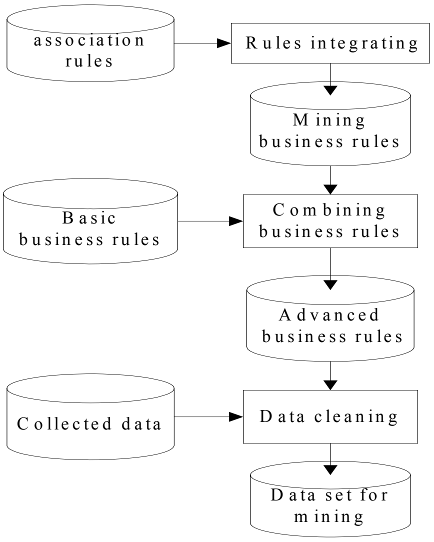

Fig.1: The basic framework of the DCAR

2.2. Definition

Let a be an attribute, v is the value of the attribute, then a=v is an item, denoted as i. And Attr(i)=a,Value(i)=v.

For example, (age=3) is an item, and denoted as i_1. Attr(i_1)=age and Value(i_1)=3. And let I={ $i_1,\ldots,i_m$} be a set of items. A transaction T is a set of items, and T $\subseteq$I. A transaction database D is a set of transactions. If X$\subseteq$T, X is a transaction too, then X is a sub set of T. And the basic definitions appearing in this paper are defined as follow:

Definition 1. (association rule) an association rule is an implication

$$R(X\Rightarrow Y\text{: } C=b\%\ ,\ S=a\%).$$

Here X and Y are sub sets of I, moreover X and Y are non-empty and disjunction sets of items. C is confidence of the rule, equals to the ratio of the number of data including X and Y to the number of data including X. And S is the support of rule, equals to the ratio of the number of data include X and Y to the number of data all. Here, $0\leqslant a, b\leqslant 100$.

There are two functions on the association rule ar: Pre(ar) = X and Post(ar)=Y, denoted as first component and last component of the ar.

For example, ({fever=middle, cough=common} ⇒ {phlegm=middle}: C=56%，S=74%) is an association rule in TCM. It means that an illness with middle fever and common cough has the middle phlegm with the probability 56 percent. And 74 percent of the ill people are in the same model.

Definition 2. (restrict of attribute value) the item set describes the possible values of the attributes, denoted as VC. VC is the set of items, and VC⊆I. Here, vc_i∈VC, Attr(vc_i)=Attr($vc_{(i+1)}$).

For example, ({fever=low, fever=middle}) is one restrict of attribute value, it means that the value of fever is low or middle.

Definition 3. (business rule) the business rule is the restrict rule in business. A business rule is an implication

BR((U⇒V) : C=c% , S=d%)

Here U and V are sub sets of VC, moreover X and Y are non-empty and disjunction sets. C is the rule confidence, and S is the rule support. It describes the rules that the data must obey.

For example, ({{fever=low, fever=middle}, cough= common} ⇒ {phlegm = middle} : C=83%, S=97%) is a business rule in TCM. It means that an illness with low or middle fever and common cough has the middle phlegm with the probability 83 percent. And 97 percent of the ill people are in the same model.

There are two functions on the business rule br: Pre(br) = U and Post(br)=V, denoted as first component and the last component of the br.

In this paper, three types of business rule are used. Basic business rule, mining business rule and advanced business rule.

Definition 4. (basic business rule) the business rule offered by the experts is called basic business rule. The basic business rule is the universal for all the data.

Definition 5. (mining business rule) the business rule mined from the data is called mining business rule. Every data source has its own mining business rules.

Definition 6. (advanced business rule) the business rule merging the basic business rule with the mining business rule is called advanced business rule. Every data source has the advanced business rule.

Definition 7. (rule base) Let ar(X⇒Y: C=b% , S=a%) be an association rule, then (Attr(Xi)) ⇒{Attr(Yj)} is called rule base of ar, denoted as RB(r). In RB(r), Xi ∈X, Yj∈Y. The rule base describes the relate attributes of an association rule. If two association rules R1 and R2 have the same rule base, it denoted as R1$=_{BA}$R2.

For example, the rule base of the association rule ({fever=middle, cough= common} ⇒ {phlegm = middle}: C=56%, S=74%) is {fever, cough} ⇒ {phlegm}

3. Generation of advanced business rules

3.1. Integrate the association rules into advanced business rules

Basing on the association rules mined from the sample of the data source, an algorithm integrating the association rules into advanced business rules called BRGAR is proposed in this section. The basic idea of the algorithm is:

- Combining the items of the association rules based on the rule base of the association rules into business rules.
- Calculating the confidence and support of the business rules.
- Merging the basic business rules with the mining business rules into advanced business rules.

Proposition 1. Assume r_1({a_1=$v_{(r1)1}$,a_2=$v_{(r1)2}$} ⇒{a_3=$v_{(r1)3}$}: C=b% , S=a%) and r_2({a_1=$v_{(r2)1}$, a_2=$v_{(r2)2}$} ⇒{a_3=$v_{(r2)3}$}: C=d%, S=c%) are two association rules, $v_{(ri)j}$ is the value of j_{th} attribute in the association rule ri. r1 and r2 have the same rule base. Then integrating with each other can get the mining business rule, denoted as br_1= $r_1$$+_{BA}$ r_2, the "$+_{BA}$" is the operator of integration based on the rule base. Then

- The result of integration is br_1 = ({a_1=$v_{(r1)1}$,a_1=$v_{(r2)1}$},{a_2=$v_{(r1)2}$, a_2=$v_{(r2)2}$}) ⇒ ({{a_3=$v_{(r1)3}$,a_3=$v_{(r2)3}$}}):C=((a+c)bd/(ad+bc)) , S= (a+c) %),When $v_{(r1)\,j}$≠$v_{(r2)\,j}$.
- The result of integration is br_1 = ({a_1=$v_{(r1)1}$}, {a_2=$v_{(r1)2}$})⇒ {{a_3=$v_{(r1)3}$}}): C=b%, S=a %) =r1=r2,When $v_{(r1)\,j}$=$v_{(r2)\,j}$.
- $r_1$$+_{BA}$$r_2$= $r_2$$+_{BA}$$r_1$

Assuming r_1, r_2 are the association rules mined from database. Let the number of the data in the database be M, the number of data including not only Pre (r_i) but also Post (r_i) be N_{ri}, the number of data including Pre (r_i) be K_{ri}. In r_1, S=N_{r1}/M=a%, C=N_{r1}/K_{r1}=b%. It means that N_{r1}=M*a% and K_{r1}=N_{r1}/b%=M*a/b. In a similar way, in $r_{2,}$ N_{r2}=M*c%, K_{r2}=M*c/d.

When $v_{(r1)j}$≠$v_{(r2)j}$, because one datum can only support one Pre(r_i) one time, the number of the data satisfying Pre (br_1) and Post (br_1) both is the sum of the N_{r1} and N_{r2} . The sum is M*(a%+c %), and then the number of data satisfying Pre (br_1) is the sum of K_{r1} and K_{r2}. The sum is M*(a/b+c/d). So S of br_1 is N_{br1}/M=(a+c)%, and the C of br_1 is N_{br1}/ K_{br1}= M*(a%+c%)/M*(a/b+c/d)=((a+c)bd/(ad+bc))%.

When $v_{(r1)\,j}$=$v_{(r2)\,j,}$ the data supporting Pre (r_1) are the data supporting Pre (r_2), so the number of data satisfying Pre (br_1) and Post (br_1) both is N_{r1} or N_{r2}, here N_{r1} = N_{r2} =M*a%. And then the number of data satisfying Pre (br_1) is K_{r1} or K_{r2}, it is M*a/b. So S of

br_1 is $N_{br1}/M=a\%$, and the C of br_1 is $N_{br1}/ K_{br1}= M*a\%/M*a/b=b\%$. Thus, $br_1= r_1$.

In the equation 3, the left of the equation is

$r_1+_{BA}r_2=(\{a_1=v_{(r1)1},a_1=v_{(r2)1}\},\{a_2=v_{(r1)2},\ a_2=v_{(r2)2}\}) \Rightarrow \{\{a_3=v_{(r1)3},\ a_3=v_{(r2)3}\}\}$: $C=((a+c)bd/(ad+bc))\%$, $S=(a+c)\%)$

The right of the equation is:

$r_2+_{BA}r_1=(\{ a_1=v_{(r2)1}, a_1=v_{(r1)1}\},\{ a_2=v_{(r2)2}, a_2=v_{(r1)2} \}) \Rightarrow \{\{a_3=v_{(r2)3},\ a_3=v_{(r1)3}\}\}$: $C=((c+a)db/(da+cb))\%$,$S=(c+a)\%)=(\{a_1=v_{(r1)1},a_1=v_{(r2)1}\},\{a_2=v_{(r1)2},\ a_2=v_{(r2)2}\}) \Rightarrow \{\{a_3=v_{(r1)3},\ a_3=v_{(r2)3}\}\}$: $C=((a+c)bd/(ad+bc))\%$,$S=(a+c)\%)$

We can see that the left of the equation equals to the right. Thus, $r_1+_{BA}r_2= r_2+_{BA}r_1$.

3.2. Advanced business rules generation algorithm based on association rules

The algorithm generating the advanced business rules based on the proposition in the last section is introduced in this section.

Suppose that, ar is the one association rule mined from the data; rs is the set of the ar; the number of ar in the rs is N; abr is the advanced business rule; abrs is the set of abr; bbr is the basic business rule; bbrs is the set of bbr; mbr is the mining business rule; mbrs is the set of mbr. Assuming the entity of the support is θ, the advanced business rules generation algorithm based on association rules (BRGAR) is described as follow:

```
Algorithm 1 BRGAR
Input：  rs, θ, bbrs
Output：abrs
Begin
    mbrs=∅
    For i=1 to N-1
        $mbr_i=ar_i$
        For j= 1 to N
            If  $ar_i=_{BA} ar_j$ Then $mbr_i= mbr_i+_{BA} ar_j$
        EndFor
    If $S(mbr_i)>θ$ and $mbr_i$ not in mbrs then
        mbrs=mbrs∪{ $mbr_i$ }
    EndFor
        abrs=bbrs∩mbrs
    Return abrs
End
```

Here, S (mbr_i) is the support of the i_{th} mining business rule. And if the rule bases of two association rules are the same, the two can be integrated. Regard it as the mbr if the support of the mbr is no less than θ.

4. The data cleaning method based on association rules

Basing on the discussion above, a data cleaning method based on association rules (DCAR) is proposed in this section. In the method, we integrate the association rules into mining business rules first; and then, based on these mining business rules, we adjust the basic business rules available by the experts to get the advanced business rules for every data source individually; At last, using the advanced business rules in appointed data source, we clean the data source. If one datum supports all the advanced business rules, the datum is a correct one; otherwise, is incorrect one.

Functionally, the method can be divided into two steps:

- Generation of advanced business rules: by invoking BRGAR, we integrate the association rules, get the mining business rules, merge with the basic business rules and get the advanced business rules.
- Cleaning the data: If a datum supports the advanced business rules all, the datum is correct one. Otherwise, the datum is incorrect even though it conflicts with one advanced business rule only.

Let d denote the database, and rs denote the association rule set mined from the sample of the one data source. Assuming the entity of the support is θ, the data cleaning method based on association rules (DCAR) is described as follow:

```
Algorithm 2 DCAR
Input: $d =\{d_1,d_2,…,d_n\},rs=\{r_1,r_2,…,r_m\}$,θ,bbrs
Output: correct data set cd
Begin
Business Rules Generation
    abrs=BRGAR (rs, θ,bbrs) ；
Data Cleaning
    ncd=∅
    cd=∅
    For i=1 to n-1
    flag=0
        For j= 1 to sizeof(abrs)
            If  IS($d_i$，RB($abrs_j$))⊄abrs$_j$   then
                ncd=ncd∩{$d_i$}
                flag=flag+1
              break
        EndFor
        if flag<>1 then cd=cd∩  {di}
    EndFor
    Return cd
End
```

Here, sizeof (abrs) is the function indicates the number of the abr in the abrs. And $abrs_j$ is j_{th} br in abrs. RB($abrs_j$) is the rule base of the $abrs_j$ IS(d_i, RB(abrs$_j$)) is the function indicates the value of attribute the i$_{th}$ datum on the attribute of rule base abrs$_j$;

5. Experiment

In this section, we present our performance study over data collected from the real world. We report our experimental results on the performance of DCAR in comparison with the algorithm DQMWAR [7], and the method, which cleans the data using the business rules offered by experts.

The main purpose of this experiment is to demonstrate how effectively to clean the data using the business rules set incorporated from the association rules mined from the data. We compare the accuracy of the methods mentioned above.

In this experiment, we need three parts of information: data sets, basic business rules and association rules.

This experiment bases on the business rules made by the experts according to the data and their experience. The standard data sets or the data sets generated by the tools cannot be used in this experiment for it is impossible for the experts to make business rules for these data sets. The data we used in the experiment are extracted from the CRF (The Case Report Form is an important tool used in clinical trials. It captures a required record of data and other information for each patient during a clinical trial as defined by the clinical protocol.) of 2 hospitals' patients in different cities, one city is Shenyang, and the other is Guangzhou. The data describe the degrees of the attributes (symptom) of the patients. The experts of TCM make the basic business rules. And the association rules are mined from the sample of the data. Every city has its own association rules. Some of the data, the degrees of symptom, and the basic business rules, association rules are showed in the Table 1-4.

The experiments were performed on a Pentium IV PC at 1.5GHz with 512MB of memory; all experiments were performed on a Windows XP machine. All algorithms were coded in Visual C++6.0. There were 10000 data from the 2 hospitals. The 6000 data come from Shenyang and the rest come from Guangzhou.

symptom	degree
fever	no fever, low fever, middle fever, high fever
cough	occasional, common, frequent, tempestuous
phlegm	no phlegm, fewer phlegm, middle phlegm, plenty phlegm
breathe	smooth, short, scant, labored
sound of pulmonary	no rale, fewer rale, middle rale, dense rale

Table 1: Symptom.

ID	fever	cough	phlem	breate	sound of pulmonay
01	low fever	frequent	no	short	fewer
02	middle fever	occasion al	fewer	short	dense
03	high fever	frequent	fewer	scant	fewer
04	high fever	tempest uous	plenty	scant	middle

Table 2: Data.

Basic business rules
{ fever=low, fever=middle ,fever=high}⇒{ phlegm=fewer }

Table 3: Basic Business Rules.

association rule
{fever=high, cough=common } ⇒ { phlegm=fewer }:C=45%,S=24%
{ fever=middle, cough=frequent } ⇒{ phlegm=plenty }: C=85%,S=22%
{ breathe=scant, cough=common } ⇒{ sound of pulmonary=ewer }:C=90%,S=65%
{ breathe=scant, cough=frequent } ⇒{ sound of pulmonary=dense }:C=63%,S=31%

Table 4: Association Rules.

In this performance test, we focus on the efficiency of the methods mentioned above. Our experiments show that in most cases, DCAR outperforms the other two. (In the Fig.2, the DMQ represents the DMQWAR, and the BG represents the method using the business rules available by the experts. The threshold of τ in the DMQWAR is 75%).

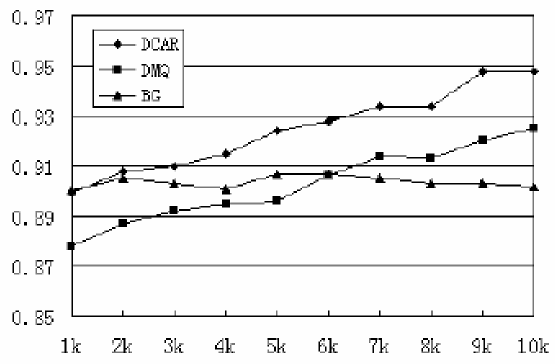

Fig.2: Accuracy.

Fig.2 demonstrates the results of performance test using the data set mentioned above. DCAR outperforms the method using the business rules available by the experts, and the performance difference becomes larger when the number of the data in dataset increases. DCAR also outperforms DQMWAR, and the performance difference becomes smaller when the number of the data in dataset increases. If the number of the data in the dataset keeps on increasing, the accuracy of the DQMWAR will increase, but no more than the accuracy of DCAR.

6. Conclusions

Data cleaning is the key step of the data-mining task, cleaning the data with the business rules made by the experts is an efficient method. Once restricts of the data are different on the different data sources, it is impossible for the experts to make the business rules. In this paper, an integration algorithm BRGAR is proposed. This algorithm integrates the association rules mined from the data sample of the data source into the mining business rules, merges the mining business rules with the basic business rules into the advanced business rules. Basing on the advanced business rules, the data cleaning method based on association rules (DCAR) is proposed. Comparing with the two algorithms, the method is beyond the two in the accuracy.

Acknowledgement

This work is partially supported by National key Technologies R&D Program in the 10th Five-year Plan of China (Grant No. 2004BA721A05).

References

[1] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, *Data Mining to Knowledge Discovery:An Overview,Advances in Knowledge Discovery and Data Mining*, AAAI Press, California, 1996.

[2] M.A. Hernandez, S.J. Stolfo, Real-World Data is Dirty: Data Cleaning and the Merge/Purge Problem. *Data mining and Knowledge Discovery*, 2:9-37,1998.

[3] F.X. Yang, Y.C. Liu, Z.H. Duan, Overview of Data Cleaning. *Application Research of Computer*, 29: 3-5, 2002.

[4] M.L. Lee, T.W. Ling, W.L. Low, IntelliClean: a Knowledge-Based Intelligent Data Cleaner. *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 290-294, 2000.

[5] F. Caruso, M. Cochinwala, U. Ganapathy, et.al., Telcordia's Database Reconciliation and Data Quality Analysis Tool. *Proceedings of the 26th International Conference on Very Large Data Bases*, pp. 615-618, 2000.

[6] B. Wang, M.W. Zhang, B. Zhang and W.J.Wei, An Effective Hypergraph Clustering in Multi-Stage Data Mining of Traditional Chinese Medicine Syndrome Differentiation. *Proceedings of the 6th IEEE International Conference on Data Mining*, pp. 332-336, 2006.

[7] J. Hipp, U. Guntzer, U.Grimmer, Data Quality Mining: Making a Virtue of Necessity, Technical Report, Workshop on Research Issues in Data Mining and Knowledge Discovery. Santa Barbara, 2001.