J. Basic. Appl. Sci. Res., 2(10)9843-9850, 2012
© 2012, TextRoad Publication

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

Fuzzy Data Mining Model Based on Quality Evaluation

Jaafar Partabian*[1], Adel Jahanbani[2], Abdulhamid Khosravi[3]

[1,2] Department of Computer Engineering, Islamic Azad University, Lamerd Branch, Lamerd, Iran
[3] Department of Economics, Payame Noor University, M.A. in Economics, Faculty Member

ABSTRACT

Regarding to the use of accurate information in informative systems, measuring the quality of information is very important in the recent years a variety of ways have been used for calculating the quality of data which mostly are based on the methods of data mining and statistic. On this research a new way of measurement of data quality is delivered.

The framework given along with the data mining method extract such a knowledge that with its help and along with a designated fuzzy system one can deliver a way for taking the measure of data quality . in this way we use two independent criteria's for calculating the quality of entering record that the first way is calculating the quality of entering record than existed rules data base and the second is calculating the quality of entering record than the proportion of each of data base fields . Then the designed system blend with the two ways delineates the final quality of entering record. the proposed method has examined on 3000 different records and in three aspects, which are the random and automatic finding of other related functions from the base time and a less consuming memory and a less mistake in calculation in compare with other methods has gained some better results.

KEY WORDS: the quality of data base**-** data mining**-** fuzzy system

1 – PREFACE

Generally informative systems nowadays are a portrayal of real world inside the computers. Members of an organization can make some products or make some decisions. With such a view on informative systems the importance of as ass the most important elements of informative systems come to view more than ever if the data of systems are not commensurate with real, world the specific organization using these systems in many of its functions such as: making correct decision will make mistake. These days' organizations are paying huge amount of sums for buying and preparing the informative systems.

But they don't pay much attention to the main basis of such systems which are data and information that gradually enter to systems and after a while they become accompany with many incorrect data and bad quality information, that they have to pay some other great sums for cleaning and increasing the quality of data [8] on a poll over the organizations with ERP that was held by LLC country group, the biggest problem on preparing and utilizing the smart commercial projects is the quality of information with is mentioned in informative systems. Also based on studies by institute of data on the matter of the quality of data annually millions of dollars are paid, [1]. However based on the interviews held by data specialists, it is estimated that between 30% to 80% percent of comments on data used for eliminating and understanding them.

With these definitions the importance of data quality will be more clarified and the proofs are the soft a wares instruments and seminars which accentuate the claims. regarding to reasons the importance and the value of management of information and controlling and measuring before the organization being damaged for existence of bad Quality data are totally proved .

The quality of data is one the most complex significances and solving the problems around the quality of data needs many information that are the result of experiments. on the other side from1990s no won that huge and great data bases were constructed, the possibility for this important that mere the experiences would help to the quality is not any more existed so the best way for adding quality can be supervising and permanent use of op statistical algorithm .

Data mining and seeking the data can create a greater scope of view and fill the data splits but the matter of automation is very fancy and the available algorithm can solve a very small part. On the quality of data 2 discussions are stated.

1**-** Managing the data quality: tries to state methods for curbing the damages. Of the most important methods, is TDQM.

***Corresponding Author:** Jaafar Partabian, Department of Computer Engineering, Islamic Azad University, Lamerd Branch, Lamerd, Iran, Email: Jaafar_partabian@yahoo.com

2– Measuring the data quality: tries on finding criteria for delineating the quality and the way of measuring.

Data quality is one of the most complicated significances. the important matter is that finding the answer of data quality needs many related info that is the result of experiences merely [1] specialists can determine the correct rule and providing such rules is a basic on evaluation of the data [3] . Date quality is a consistent accomplish that is continued from the start to the end of the stage, is continued. The request for updating the definitions is sensed more and more every day. And that's why the process of data quality and needed indexes for measuring are scrutinized.

2-Symols and definitions

In this chapter we will work on definitions of symbols and expressions which are needed.

D: Data base

R: entering record that is consist of a collection of given info which are needed to be assessed on their quality.

Min _ supp (least of percentage of support):

That's a data from a database which is supported by the base and its amount is determined by user.

Min _ conf :(least of percentage of confidence):

It is the times that if the A data is occurred then the B data is appeared and its amount is determined by user.

Conf (A→B) = $\dfrac{SUPP(A \rightarrow B)}{SUPP(A)}$

Functional dependency:

If A and B are two volunteer I tens of data base we say, B is depend on A and is depicted with A→B which the amount of A determines the amount of B. just if in every possible amount of data base for every A we have an amount of B.

Dependency rule:

If a and b are two of data base in this if xεA and yεB and then x→y a dependency x is called the body and y is called the top on dependency rule[2] .

Business Rules:

Dependency rules extracted from base which are the main for evolution of quality entering record than the previous data of base and are shown with BR are put.

Homogeneous and in homogeneous rules:

If the extracted dependency rule of entrance record is fit with one of the extracted business rules, we call homogeneous rule and otherwise if only the body of extracted rule of entrance record is fit with one of the business rules we call it un adjustable rule.

Strange rule :

The extracted rule from entrance record which is not adjustable or un adjustable or on the other word extracted dependency rule from entrance record which has a trivia confident percentage.

Adjustable and un adjustable feat use:

The amount of each of field of the entrance record which has been examined in related on field data base is adjustable feature, and otherwise is un adjustable feature.

3-Suggestive way

The aim of this research is evaluation of quality of entrance record to the previous data base on the other word, it is tried with the use of data mining technique be evaluated for assessing the quality of entrance record we was two criteria .

First critera *:* the quality of entrance record other than BR (Business Rules) which is extracted from data base.

Second critrias*:* the quality of entrance record rather than the amount of each of data base fields (DR).

In the end by using of these two criteria and designed fuzzy .

System the final quantity of the entrance record quality is delineated.

3-1 Extract of a collection of business rules from data base :

In this way with some modifications on **APRIORI** data mining algorithm have been exerted [5]. We have made the algorithm capable of finding and extracting all functional dependency in the data base. in this method in the first pass, we extract all of given info of every table field which their numbers are higher than defined min-supp by user .

Then in the second pass two-two relationship between all of extracted given information in first pass, will be scrutinized and we achieve the amount of confidence in every given information.

if the amount of gained confidence is higher that the amount of defined min-conf, we select it as a dependence rule and we add to the collection of dependence rules, which are extracted. so in the end of second pass all of dependence rules are extracted . in the third pass regarding to extraction of all of dependence rules in previous pass, we discover the functional dependences existed among the data base information and then we extract the dependence rules fit to all functional dependency and collect them as a collection of business rules that as the accurate rules

Are the basis of evaluation of entrance record rather than the info of data base which quality is used .

Example: D, data base is supposed with the table of clerk .

Table of clerk consists of following info.

Name, Zip code, State, City, job, Salary

Min_ supp=%10 , min _ conf=0.7 : in this example , the collection of achieved numbers from the first pass, an as following .

Name	**Salary**	**Job**
ali **adel** **jaafar** **hassan**	20 5 10 15	Manager Guarder door keeper Expert
ZipCode	City	State
831 **830**	lamerd lar assaloye kangan	fars boshehr

Table 1

Some of the qehieued dependency rules from D data base , are as following.

Manager	→	**20**
Guarder	→	**5**
door keeper	→	**10**
Expert	→	**15**
assaloye	→	**830**
assaloye	→	**boshehr**
lar	→	**fars**
kangan	→	**boshehr**
Lamerd	→	**831**
fars	→	**831**
boshehr	→	**830**
lamerd	→	**fars**

Figure 2

In the third puss regarding to dependence rules in the past puss , collection of functional dependence is extracted like following:

Job	→	**salary**
City	→	**zipcode**
City	→	**state**
State	→	**zipcode**

Table 3

At the end all of dependence rules fit to. Functional dependence, which are called business rules are extracted.

Here you can see some parts of business Rules rules as following :

Manager → 20	
Guard → 5	
doorkeeper → 10	
Expert → 15	
assaloye → 830	
assaloye → boshehr	
lar → fars	
kangan → boshehr	
lamerd → 831	
fars → 831	
boshehr → 831	
lamerd → fars	

Table 4

By a collection of business rulers gained from D data base we use as basis of evaluation of entrance record.

3-2- Extraction of a collection of homogeneous, in homogeneous rules and strange rules of entrance record

For calculating the quality of entrance record to the previous Data , a data base is necessary , also a collection of adjustable rules unadjustable rules and strange rules existed between the entrance record of fields are discovered and based on this collections we can evaluate the quality

Algorithm (1)

1-Try to achieve the collection of dependence rules which is between entrance record that is fit on extracted functional dependence from third pass.
2- If each of dependence rules is commensurate with one of the rules of the business rules collection, add to the collection of adjustable rules. Other wise, each of dependence rules of record like x→y that X is the body of rule and Y is dependence rule, if X is existed in the body rule of business and Y is not existed in business rules, add it to the unadjustable rules.
3- And if the body of dependence that is achieved of record is not existed on the collection of business rules , add if to strange rule .

At the end of this stage, some collections of adjustable rules un adjustable rule and strange rules of entrance record have been discovered. as according to algorithm 1 , each of the dependence rule of the record in one of the mentioned collections has a role and it is necessary to say that in the next stage , these 3 collections will be used for calculation of the amount of the quality of entrance record .

3-3- the function of calculation of BR

Her calculation , we use the following diagram

$$BR = \frac{\text{Number of homogeneous collection element}}{\text{Number of homogeneous collection element + number of in homogenous collection element}} *100$$

In continuation of previous entrance record if it is supposed , then , the collection of adjustable rules , an adjustable rules and strange rules equal to entrance record is as following :

zipcode	City	State	Job	Salary	Name
830	lamerd	fars	manager	20	adel

Collection of strange rules	Collection of inhomogeneous rules	Collection of homogeneous rules
Director manager → 20	fars → 830 lamerd → 830	lamerd → fars

As, the rule of FARS →LAMERD is equal to one of the rules of business in table 4, so it is one the adjustable rules . on the other side , because of 830 →FARS and 830→LAMERD just their bodies is comonensurate with table 4, it is an homogeneous rule.20→ Director manager is an strange rule . Because neither is it commensurate with the business rules of table 4,por is its body commensurate with. The amount of BR for the following is equal to :

$$BR = \frac{\text{Number of homogeneous collection element}}{\text{Number of homogeneous collection element + number of in homogenous collection element}} *100$$

$$\mathbf{BR=\frac{1}{3}*100=33\%}$$

3- 4 calculation of DR Criteria

This criteria is a manifestation of quality of entrance record to the amount of every of the data base fields which is independent from the BR criteria to evaluation of quality of en trance record. For calculating the DR criteria , the second algorithm id mentioned .

Algotithm2

1-The amount of each entrance record field should be extracted and be compared to the other amount, if the base had experienced that special amount , add the amount to the collection of homogeneous . other wise , add this amount to the collection of in homogeneous.

2-Regarding to the following equation we find the amount of DR.

$$DR = \frac{\text{Number of homogeneous collection element}}{\text{Number of homogeneous collection element + number of in homogenous collection element}} *100$$

For example for record of section 3-3 of table 5,is achieved as collection of adjustable and an adjustable features of D data base.

Inhomogeneous featuer	**homogeneous featuer**
Director manager	**lamerd fars**

Table 5

$$DR = \frac{\text{Number of homogeneous collection element}}{\text{Number of homogeneous collection element + number of in homogenous collection element}} *100$$

3-5 : calculation of final quality of entrance record

As what was shown on the previous chapters, two amount were gained based on DR .and BR . that the, amount of BR which is gained , is always greater or equal to the real quality , because about the strange rules , there is no statement . for mil ding the gained amount of BR , we describe DR criteria which was told previously. As the quality has a phase identity , for finding a unit amount for showing the final quality of entrance record which is found on the mixing of gained amount from BR . and DR criteria , we act according to algorithm 3.

(algorithm3)

If DR= loo% or BR =0 them Q= BR
(Q= the final quality of entrance record)
Other wise , the final record will be mentioned by designed phase system which its entrance
is consist of DR and BR .

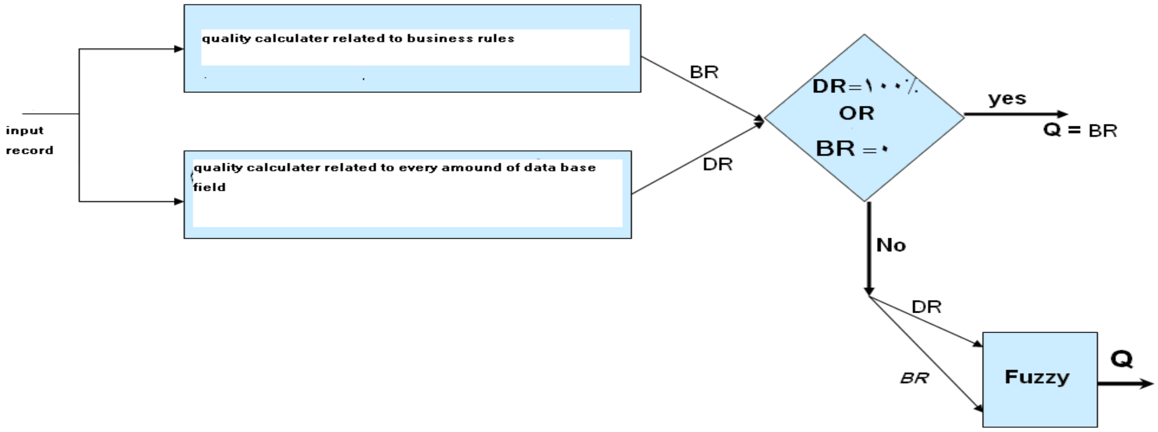

Figure 1

As what was mentioned , algorithm 3 uses the phase system for calculating the quality of entrance record

3-6 designing a phase system

For designing a phase system , we need a group of rules . then in trod use to understand system and next , according to the functions of membership and a special combination of rules , united result can be taken. The typical and traditional algorithms which are used for designing the phase system , mostly use Mamdni method .

In the inferring phase the method used , most of outputs use the centrifugal method .

Phase rules which are defined for the designing the phase system in the algorithm 3, have been defined as following :

If (BR=low) and (DR=high) $\longrightarrow$ Q = low
If (BR= high) and (DR= high) $\longrightarrow$ Q = medium
If (BR= low) and (DR= low) $\longrightarrow$ Q = low
If (BR=high) and (DR= low) $\longrightarrow$ Q = low

3-6-1 :the designed membership funetion.

The specific amount in the DR ,BR , Q membership functions are calculated as experiment and reaction and it is tried to consider the membership functions which show the nearest amount to the real.

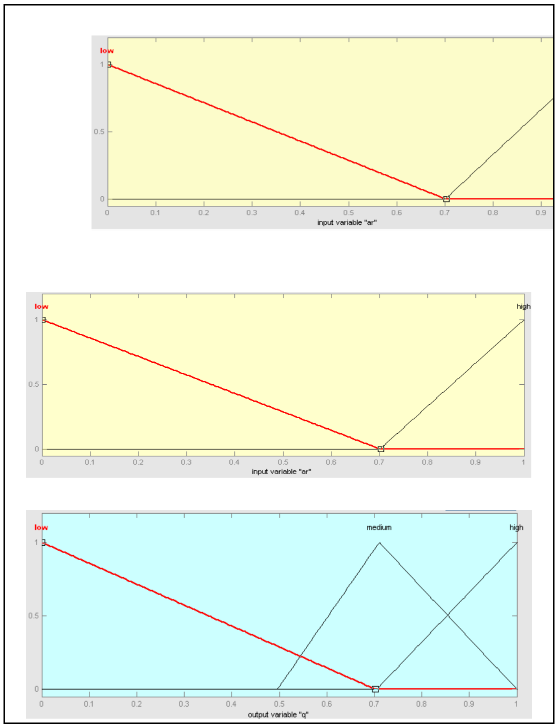

Figure 2

In the end the designed phase system . shows a Scaler number as the quality of entrance record .

4: calculation of the quality of data base .

Regarding to this fact that inside data of informative systems enter by entrance records, we can say . that data base is a group of records. So, the quality of data base can be found from the quality of records. For this , we use algorithm below .

Algorithm4

1- Calculation of the quality of all entrance records based on algorithm3.
(the quality of entrance records can be restored during the time in a place)
2- Calculation of the average quality of existed records as following aquation.

$$\text{The quality of data base} = \frac{\sum_{i=1}^{n} Qi}{n} * 100$$

the gained amount from algorithm 4 , can clearly show the quality of data base .

analyzing the method and comparing to Hipp method

1- finding the automatic functional dependence. this method randomly can find and extract the functional dependences of data base and decreases the users interfering .

2- Less time and consuming memory .

Because , the proposed method does not need to extract all dependence rules of data base . (contrary with HIPP and peers)(6)

So, the less time and memory are needed for quality evaluation of data . (no.3)

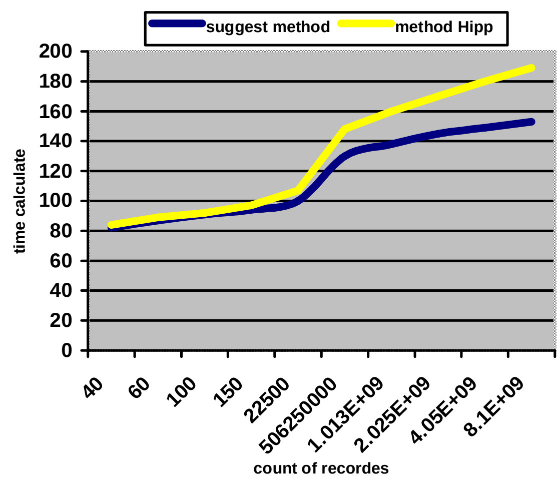

Figure 3

2- The fewer mistake on the quality evaluation of entrance record .

We exert the proposed method on 3000 data base , after the quality evaluation of data base with the proposal method and comparing with the real quality of data base (real quality is that the expert one is able to calculate base on his knowledge). It was observed that 71% of all is fit with the real quality . but in the HIPP method , it was observed that about 50% of all was need to the real result .

5-conclusion

In this article we proposed a new way for calculating the quality of data base which two relative quality were calculated for each entrance record that according to these two,and delivered algorithms , the final quality of records was gained with the use of phase system . Then by taking the average of the quality of entrance records , we could calculate the quality.

REFRENSES

[1]Haghi abdolhamid,” Data maining in Data qulity”,Iran center magazine,year 17,no 2,summer 2006.

[2]Rohani rankohi,”Data base”,edit 4,2004.

[3] Allen,S.,”Name and Address Data quality “,computerworld,(2002).

[4]Bobrowski,m.,Maree,M.,Yankelevich,D.,”Homogenueous Framwork to Measure Data Quality”, *Communications of the ACM,55,13,(2003).*

[5] Christian Borgelt, Intruduction of Association Rules apriori hmplementation. Depeartmant of knowledge processing and language engineering.oct 2003

[6] Hipp,J.,G¨untzer,U.,Grimmer,U.,”Data Quality Mining”, 3rd International

Conference on Practical Aspects of Knowledge Management,(2003).

[7] Loshin, D.,“Rule-Based Data Quality,” *Proceedings of the Eleventh International Conference on Information and Knowledge Management,(* 2002).

[8] Strong, D. M., and Lee, Y. W., and Wang, R. Y. “Data Quality In Context,” *Communications of the ACM 40*, 5, (2004).