A DECISION RULE METHOD FOR DATA QUALITY ASSESSMENT

PROCEEDINGS

(Completed Paper)

Nawaf Alkharboush
Computer Science Discipline, Faculty of Science and Technology
Queensland University of Technology, QLD 4000, Australia
n.alkaharbosh@student.qut.edu.au

Yuefeng Li
Computer Science Discipline, Faculty of Science and Technology
Queensland University of Technology, QLD 4000, Australia
y2.li@qut.edu.au

Abstract: The assessment of data quality is a key success factor for organisational performance. It supports managers and executives to clearly identify and reveal defective data in their information systems, and consequently minimises and eliminates the risks associated with decisions based on poor data. Despite the importance of data quality assessment, limited research has been conducted on providing an objective data quality assessment. Researchers and practitioners usually rely on an error ratio metric to calculate abnormal data. However, this approach is insufficient in terms of providing a complete quality assessment since errors can be randomly and systematically distributed across databases. This study will introduce a decision rule method for providing a comprehensive quality assessment, which captures and allocates quality change at the early stage in organisational information systems. A decision rule can also be extended to answer important questions such as the randomness degree and the probability distribution of errors. These advantages will significantly reduce the time and costs associated with performing quality assessment tasks. More importantly, the efficiency and effectiveness of the decision rule for assessing data quality enables management to make accurate decisions reflecting positively on organizational values.

Key Words: Data Quality Assessment, Decision Rules, Information Quality, Data Mining

1. INTRODUCTION

Data quality has increasingly become a critical concern for organisations. Rapid growth in the size and technologies in databases and data warehouses of organisations have bought significant advantages. Managers and employees can easily store massive amount of information, retrieve information, find valuable customers, and predict future opportunities and risks, and these benefits are all available when the quality of the information systems is high. However, in the real world, a

database or data warehouse is usually impacted by the appearance of error values. It is well known that errors are systematically and randomly pervasive across a database and data warehouse causing poor data quality, which can have a severe impact on customer satisfaction, decision making, operational cost, and strategy execution. A study has shown that the estimated immediate cost stemming from the 1-5% error rate is approximately 10% of revenue [13]. Additionally, poor data quality can reduce the credibility and useability of the databases and data warehouses.

Data quality can be broadly defined as data fit for use by data consumers [1]. Specifically, data quality is defined as a multi-dimensional concept, and the dimensions of accuracy, currency, completeness and consistency are regularly mentioned [1,14,15]. These dimensions are most pertinent to data values. Accuracy refers to the value that is nearest to the value in the standard domain, and the new datum will be compared to standard domain datum which is considered as accurate (or correct) in determining the accuracy of a new datum. Currency or (timeliness) means that data value is up-to-date, and this dimension can influence the accuracy of decisions. For instance, people usually receive mail posts from commercial and governmental sectors which do not belong to us and belong to the previous resident. The third dimension is completeness, which refers to the absence of missing (or unknown) values in a data collection. The finally dimension is consistency, which means that there is no conflict among data values. Conversely, inconsistency (or outliers in the data mining context) presents values that are out of range of the remaining data collection. Grouping quality dimensions into these four categories is critical for conducting appropriate quality assessment and research. The literature uses the term 'data quality' synonymously with information quality.

In the data quality literature, it is accepted that the assessment and improvement of organisational information systems cannot be achieved independently from data consumers (or user perspective) who use the data. Indeed, data consumers play significant roles in providing a comprehensive and continuous quality assessment and improvement. In the manufacturing situation, for example, producing accepted products that meet customers' needs is highly dependent upon the quality of the inter-process. Similarly, in the context of information quality, a clear business process can improve the quality of data stored in databases and data warehouses.

Data consumers decide whether or not data values are error free and appropriate for their needs. Several outstanding theoretical methodologies have been proposed to investigate the quality assessment from the users' perspective [3,6,14], and these methods mainly depend on qualitative studies. However, the outcomes of these studies prove that, within an organisation, different departments can provide different quality assessment for data quality dimensions accuracy, timeliness, completeness and consistency [11]. Another shortcoming is that subjective assessment consumes much labour, effort and time, and requires manual inspection of data values by users to determine the quality of data values, therefore increasing the cost associated with quality assessment tasks. Therefore, it would be beneficial for data consumers to employ an efficient objective assessment that objectively reports the quality status of a database and the location of abnormal data values.

Quality of data can also be assessed from the data perspective, and this view is concerned with those four quality dimensions that pertain to the data themselves. These dimensions include accuracy, currency, completeness, and consistency and can be revealed statistically. However, much research is needed to improve the objective quality assessment from the data perspective. The most common measurement for quality assessment purposes relies on the error rate to calculate the ratio between the number of correct values and the total number of values for targeted databases or data warehouses [1,2,3,11,12,16]. This approach is insufficient for providing managers with valuable information such as benchmarking quality change from the previous year to the current year, allocating the location of defective data, and estimating the time and costs needed for quality improvement.

Additionally, managers cannot apply the error rate method when comparing the quality issues across databases because errors are randomly and systematically distributed across columns and rows. A recent study [5] extends this error rate approach to include randomness and probability distribution to enhance the accuracy assessment of a database. However, as the authors indicate, their approach has time complexity which makes it impractical for a large database.

This study will extend the current error rate metric to include the decision rules method for providing complete and accurate quality assessment. A decision rule efficiently and effectively identifies quality issues in organisational information systems. Managers and executives can rely on the proposed method to conduct a complete quality assessment for their information resources, which assists decision makers to quickly define any quality change in either positive or negative ways. Additionally, the proposed method specifies the degree of systematic or random errors and their probability. This will benefit management to estimate the degree of errors in a database and therefore accurately estimate the time and costs required to improve the quality. It will also allow executives to reduce or eliminate the risk and impact of poor data on an organisation's performance as well as increase satisfaction among stockholders. The main contributions of this research include: assessing the quality of single or multiple databases, defining quality change across databases, allocating abnormal data values, measuring the systematic and randomness degree of errors, and calculating the probability distribution of the errors.

This paper is organised as follows: Section 2, analyses the related literature, Section 3 defines the proposed model, Section 4 describes empirical experiments, Section 5 analyses the results, and Section 6 provides a conclusion and possible future direction for research.

2. RELEVANT BACKGROUND

2.1 REASONS FOR ASSESSMENT

Data quality involves a series of steps including defining, measuring, analysing, and improving data quality. These four components are gradually and seamlessly interrelated to each other, and are required for a comprehensive and good quality program. For instance, management cannot improve the systems if they cannot define or measure defective values. Similarly, improving quality of data in organisations' systems without improving the process will not ensure continuous improvement. Deficient values will continue to be stored, resulting in incomplete, inconsistent, and inaccurate information in organizational information systems. This process is advocated by leading data quality programs such as MIT Total Data Quality Management (TDQM) and Department of Defence (DoD) [7].

Researchers and practitioners have investigated these four components and provided outstanding theoretical and practical contributions. For the measurement phase, studies have introduced two views which include the user perspective and data perspective for quality assessments. Both views are intrinsic for benchmarks for the processes and databases to improve quality [12]. Assessment of data quality is the first step toward delivering high data quality, which enables decision makers and management to track down the areas that produce quality conflict. Without assessment of the process and data quality, it is difficult to anticipate what and where effort should be made to improve the data quality. Ballou and Pazer said, that "which does not get measured does not get managed" [1]. The awareness of the quality level of a database or a data warehouse enables management to capture the root causes of the quality problems, reduce the impact of defective data and create a roadmap for achieving quality improvement.

Quality assessment of the processes and databases needs to be measured to determine if there is degradation or improvement of the degree of data quality. This allows management to measure the current quality and estimate anticipated quality improvement tasks. Users can reduce and eliminate the risk of making incorrect decisions that could have severe impacts on customers' satisfaction. Further, by benchmarking the process and database, decision makers can clearly and specifically allocate the cause of quality. Quality assessment is significantly useful in a situation where managers cannot or do not have the resources for improving the data quality, since quality assessment can estimate the costs or risk of the decisions made based on incorrect data quality.

2.2 CURRENT QUALITY ASSESSMENT

The most common measurement for data quality is based on calculating the error ratio or accuracy ratio. After deficient data values have been disclosed, users can easily employ the error rate method by counting the total of defective data fields divided by the total fields of table [2,4,11,12]. Alternatively, users can calculate the accuracy rate by subtracting 1 from the result of error rate "accuracy rating = 1- (total numbers of defective data/total numbers of fields)". This method is useful for presenting the ratio of wrong or correct data values. However, such assessment is insufficient and offers inefficient information to managers to improve data quality. Also, the error rate method might report the same error or accuracy ratio of databases, which can be misleading since errors are randomly and systematically distributed across databases. For instance, if management want to assess the quality of three databases or the quality of data over the last three years, the error rate method might report the same results. This might be incorrect because errors might be distributed in different locations: for example, the first database might be systematically distributed in one column and systematically distributed in one row for the second database, and randomly across columns and rows for the third database.

A recent study [5] proposes a new quality assessment method. The authors extend the current error rate method to include a randomness measure and probability distribution to enhance the quality assessment from data point of view. In their study, quality measurement is based on the three vectors including error rate, randomness measure and probability distribution. Firstly, they calculate error rate based on the same description as in the previous paragraph. Then, the authors adopt the Lempel-Ziv (LZ) complexity algorithm in order to determine whether the errors in databases are randomly or systematically distributed. Finally, the study adopts the Poisson distribution method to measure the probability of the errors in a database, due to the fact that some errors are higher in some rows and columns than others. However, this study is inefficient for assessing a large database. The problem as the authors indicate is that the Lempel-Ziv (LZ) algorithm has a time complexity, which makes it inadequate for assessing a large database. Therefore, it is suggested that a database has to be randomly segmented into small samples to compute the LZ algorithm. In this case, users have to run the LZ algorithm several times, which is impractical.

This study aims to propose a complete quality assessment method. The proposed decision rules method will provide management with accurate and valuable information to improve data quality, so they can clearly determine any positive or negative quality change across databases or over time. They also can clearly specify whether errors are randomly or systematically distributed across a database. Furthermore, decision makers can easily and efficiently determine the probability distribution of errors in a database. These advantages will enhance the accuracy of quality assessment and therefore enhance the accuracy of the decision making, and provide management with information required to accurately estimate data quality improvement**.**

3. THE DECISION RULE METHOD

3.1 DECSISON RULE METHOD FOR QUALITY ASSESSMENT

Rough set theory formulates the foundation of our quality assessment method. It has been increasingly used in many interesting applications. In research areas such as data mining, machine learning, knowledge discovery, and decision analysis, rough set theory has shown significant contributions. Users can describe the knowledge in information tables [9, 10] or multi-tier structures [17, 18, 19, 20]. Additionally, users can represent the association among data. Despite the popularity of rough set theory, little research has been conducted in the areas of quality assessment and quality improvement.

In this paper, the proposed decision rule method is used to provide management with information needs for data quality assessment. Management will be able to determine any change in the quality across databases or years in a single database. This study assumes that there are two databases D_1 and D_2 with the same data structure, where D_1 is called a history database or a training set; and D_2 is a newly generated database or a testing set. Formally, D_1 (or D_2) can be described as multiple decision tables (G_i, A_i), where G_i is a set of granules about attributes A_i, and a granule is a group of objects (rows) which has the same attributes' values [10,17].

For example, Table 1 is a simple database which includes 6 rows and 7 attributes. We represent normal and defective data values as 0 and 1, respectively. Table 1 can be compressed in a decision table as in Table 2. A decision table includes only 4 granules in Table 2, where the support of granule, $sup(g_i)$, is the number of rows with the same values for the 7 attributes, also called the size of covering set of the corresponding granule.

Attributes A_i can be divided into two groups: condition attributes (C_i) and decision attributes (D_i), such that $C_i \cap D_i = \varnothing$ and $C_i \cup D_i \subseteq A_i$. Every granule in the decision table can be mapped into a decision rule, e.g., g_1 in Table 2 can be read as the following decision rule:

$$(a_1 = 1 \wedge a_2 = 1 \wedge a_3 = 0 \wedge a_4 = 0 \wedge a_5 = 0) \rightarrow (a_6 = 0 \wedge a_7 = 0)$$

or in short $C_i(g_1) \rightarrow D_i(g_1)$, if $C_i = \{a_1, \ldots, a_5\}$ and $D_i = \{a_6, a_7\}$, where ^ means "*and*" operation.

Users can assign condition attributes and decision attributes according to different requirements about the data quality problems. For example, normal rules can be the condition and defective rules can be the decision.

Table 1. A database

Row	a_1	a_2	a_3	a_4	a_5	a_6	a_7
1	1	1	0	0	0	0	0
2	0	0	1	1	0	1	0
3	0	0	1	1	1	1	0
4	1	1	0	0	0	1	1
5	0	0	1	1	1	1	0
6	1	1	0	0	0	1	1

Table 2. A decision table

Granule	a_1	a_2	a_3	a_4	a_5	a_6	a_7	$sup(g_i)$
g_1	1	1	0	0	0	0	0	1
g_2	0	0	1	1	0	1	0	1
g_3	0	0	1	1	1	1	0	2
g_4	1	1	0	0	0	1	1	2

The database can also be represented as a small decision table if the user only considers a subset of attributes. For example, Table 3 and Table 4 show two small decision tables.

Table 3. A small decision table, where $A_i = \{a_1, \ldots, a_5\}$.

Granule	a_1	a_2	a_3	a_4	a_5	$sup(cg_i)$
cg_1	1	1	0	0	0	3
cg_2	0	0	1	1	0	1
cg_3	0	0	1	1	1	2

Table 4. A small decision table, where $A_i=\{a_6, a_7\}$.

Granule	a_6	a_7	$sup(dg_i)$
dg_1	0	0	1
dg_2	1	0	3
dg_3	1	1	2

The attributes in databases are normally organized in multi-levels (or a hieratical structure), such as product categories [17, 19]. For example, attributes $a_1,\ldots, a_5$ may be in high level, i.e., Category 1 , and attributes a_6 *and* a_7 is also in high level, i.e., Category 2. Therefore, we can have the high level decision rules between Category 1 and Category 2, which have the form as follows:

$$\boldsymbol{cg_i \rightarrow dg_j}$$

For the above examples, we have the following 4 decision rules between Category1 and Category2:

$$cg_1 \rightarrow dg_1\,;\; cg_1 \rightarrow dg_3\,;\; cg_2 \rightarrow dg_2\,;\; cg_3 \rightarrow dg_2$$

where $cg_1 \wedge dg_1 = g_1$, $cg_1 \wedge dg_3 = g_4$, $cg_2 \wedge dg_2 = g_2$ and $cg_3 \wedge dg_2 = g_3$ (please see Table 2, Table 3 and Table 4).

The *support* of rule $(cg_i \rightarrow dg_j)$ is $sup(cg_i \wedge dg_j)$, where $(cg_i \wedge dg_j)$ is a granule of Table 2; and the *confidence* is

$$\frac{sup\ (cg_i \wedge dg_j)}{sup\ (cg_i)}$$

A decision rules can also be utilised to evaluate the quality of data from multiple databases. Users can assign condition attributes as D_1 *a history* database or training set and assigns decision attributers as D_2 *a newly generated* database or testing set. Decision rules can be discovered from D_1 and make matching in D_2. The result will determine if there is quality change in a new database. If there are quality problems, the number of rules in D_1 is not matched with the ones in D_2 or the support for the matched rules in D_2 is significantly higher than correspondent rules in D_1. Decision rules are also useful to determine if there has been improvement on data quality. For instance, if the numbers of rules in D_2 are low and these rules match with D_1, the quality of D_2 is therefore getting better. Another indication for quality improvement is that the ratios of support in defective rules in D_2 are less than the corresponding ones in D_1. By looking at the numbers of support, the managers can determine the frequency of the defective rules. Therefore, decision rules for data quality in multiple databases D_1 and D_2 can be defined in the following formula:

$cg_i \rightarrow dg_j$ where cg_i is a *condition* granule and dg_j is a decision granule.

In summary, the above idea can be formally described as the following two processes: the training and testing.

Training Process

(1) Scan the database, D_1, to define the normal or abnormal (defective) data values;
(2) Transform normal data value into "0" and abnormal data value into "1" in the D_1;
(3) Generate the corresponding decision table (G_1, A) from D_1 by grouping rows with the same attributes' values, where A is the selected attributes in D_1.

Testing Process

(1) Process D_2 as the same as for D_1, and assume a decision table (G_2, A) is obtained;
(2) Compare the defective rules between the decision table (G_1, A) and the decision table (G_2, A); and calculate the numbers of matched rules and unmatched rules;
(3) For the matched rules, determine the severity of quality problem in D_2 by measuring the differences of the support and confidence;

(4) Calculate the randomness degree of the defective data using the defective rules; and analyse the error distributions, and report where the possible problems are.

The details for the analysis in the testing process will be discussed in the next session.

3.2 RANDOMNESS AND ERRORS DISTRUBUTION MEASURMENT

Randomness measurement is essential for delivering a complete method for data quality assessment. Randomness can be defined as the lack of rules that govern the construction pattern [5]. In databases, errors are either systematically or randomly distributed. A systematic error across a database presents a clear pattern. On the other hand, randomness error has no underling pattern. Literature has provided various techniques and methods for handling both systematic and randomness errors. Intuitively, handling errors in systematic type is less complicated than randomness errors [5]. Randomness measurement enables management to correctly estimate the costs and time needed to improve the data quality.

Table 5 (A) Cover all possible rules

Granule	a_1	a_2	a_3	*support*	*probability*
g_1	0	0	0	50	0
g_2	0	0	1	15	0.3
g_3	0	1	1	5	0.1
g_4	0	1	0	6	0.12
g_5	1	1	1	4	0.08
g_6	1	1	0	3	0.06
g_7	1	0	0	10	0.2
g_8	1	0	1	7	0.14
…					

Table 5 (B) Not cover all possible rules

Granule	a_1	a_2	a_3	*support*	*probability*
g_1	*0*	*0*	*0*	*50*	*0*
g_2	*0*	*0*	*1*	*25*	0.5
g_3	*0*	*1*	*1*	*10*	0.2
g_4	*0*	*1*	*0*	*15*	0.3

A decision rule computes the randomness degree from the decision table. Regardless the size of a database, the size of a decision table can be calculated based on the total numbers of attributes. This enables user to predict the numbers of rules to get a complete size of a decision table. We use the term covering size to calculate the largest size of decision table as illustrated in equation (1).

$$Covering\_size = 2^{|Ai|} \quad (1)$$

where the $A_i = \{a_1, a_2 \ldots, a_m\}$, *and* $|A_i|$ is the total number of attributes. For example, in Table 5 (A), the number of granules is just the *Covering_size* that means the table includes all possible rules generated from a database, where $|A_i| = 3$; however, in Table 5 (B), the number of granules is less that the *Covering_size*. Table 5 (A) also includes 1 normal rule (g_1, which not included any defective values) and 7 defective rules, g_2 to g_8.

We can modify Eq. (1) to calculate the all possible defective rules, as illustrated in equation (2).

$$Covering\_size\ of\ Defective\ Rules = 2^{|Ai|} - 1 \quad (2)$$

Based on Eq. (2), we define the randomness degree of errors as follows:

$$Randomness\ degree = \frac{The\ numbes\ of\ actual\ defective\ rules}{2^{|Ai|}-1} \quad (3)$$

In Table 5 (A) for instance, the randomness degree in a decision table is the numbers of defective rules (which is 7 rules, g_2 to g_8) divide by covering size of defective rules (which is also 7 see Eq. (2)). Therefore, the randomness degree of decision table 5 (A) = 7/7 = 100%.

However, in some scenarios, the number of rules in a decision tables are less than the covering size. Table 5 (B) for instance, includes 1 normal rule and 3 defective rules. This indicates that errors do not occur in all possible forms in a database which means less randomness degree. The randomness degree of errors in Table 5 (B) is 3/7 ≈ 43%. Randomness measurement assists managers to determine the degree level of randomness and therefore anticipate the time and the costs required for cleaning deficient data. A high percentage means the lack of underlying pattern and hence more randomness. The inverse is also true.

Another critical component for quality assessment is measuring the probability distribution of errors. It appears in a database that some errors occur in some columns more than other. It is also common that some rows have more than one error. A decision rule can be extended to measure the probability distribution of errors. For this purpose, we firstly need to count the support of defective rules. We can also calculate the probability of each defective rule, g, using the following equation:

$$Probability(g) = \frac{sup(g)}{\sum_{g_i} sup\ (g_i)} \quad (4)$$

where $\sum_{g_i} sup\ (g_i)$ is the total support of defective rules. For example, in Table 5 (A), $\sum_{g_i} sup\ (g_i)$ = $sup(g_2)$+ $sup(g_3)$+ ... + $sup(g_8)$ = 50%. Therefore, users can easily apply Eq. (4) to calculate the probability of defective rules. By calculating the probability distribution, decision makers can confidently decide which of rules have highest and lowest severity of data quality problems and therefore, determining if their information systems are in needs for urgent quality improvement or not.

4. EMPIRICAL EXPERIMENTS

4.1 EMPIRICAL DESIGN

The experiments of this study implement on a real store database which obtains from http://sky.scitech.qut.edu.au/~li3/Granule_mining/GM_Introduction.htm. We run our experimental study on 10 attributes and 2977 rows. We impact the quality of the original database by 5%,10% and 15% in order to examine the efficiency of our model. The results obtain from four databases (original, 5%, 10% and 15%) are encouraging and proving that decision rules method is reliable for assessing the quality of organisational information systems. We divide each of four databases (original, 5%, 10% and 15%) into two part training set or D_1 which consists of 1489 rows and testing set D_2 which contains 1488 rows. D_1 is a history database and D_2 is a newly generated database, see Table 6.

Table 6. Numbers of Rules

Database	D_1	D_2
Original	*73*	*67*
5%	*277*	*286*
10%	*376*	*373*
15%	*495*	*472*

In both D_1 and D_2, we construct decision tables for all four databases (original, 5%, 10% and 15%). This will help to group all similar rows together and measure the frequency of similar rows. After compressed a transaction records of each database, we obtain the decision table which includes numbers of rules or 'granules', see Table 6. For example the numbers of rules in the original database for D_1 which has 1489 rows is 73 rules and for D_2 which has 1488 rows gets 67 rules. Constructing a decision table will significantly reduce the size of a database without losing information [20] compare with association rule mining. Then, we compare between rules in D_1 with rules in D_2 to determine then data quality problems. If there are many new defective rules in D_2 that indicates there is a quality problem in the new database. The support column in a decision table is also used to determine unmatched rules that have or have not severed quality problem.

Another critical part of the proposed method is to measure the randomness of errors. Errors are usually randomly and systematically distributed across column and rows. Systematic errors have clear pattern. Unlike systematic errors, randomness errors have no regular pattern. Dealing with random errors type is far more complicated than systematic errors [5]. Randomness errors require much time and effort in order to improve its abnormal data. Therefore, it is necessary that managements measure the randomness degree of defective data. This will assist decision makers to estimate the complexity and the time associated with random errors. A decision rules method assesses the randomness degree of errors by simply dividing defective numbers of rules to a total size or decision table.

A side of decision rules is to estimate the probability of errors distribution. It is common emerge that some rows and columns have more errors than other. Therefore, a decision rule method measures the probability of errors by dividing the support of each deficient rule to the total support of defective rules. By doing this, managers and executives can clearly determine whether or not these rules in need for urgent quality improvement.

5. RESULT AND DISCUSSION

Table 7. Rate of Match and Unmatched Rules

Database	D_1 *Rules*	D_2 *Rules*	*Rule Match (%)*	*Unmatched Rules(%)*
Original	*73*	*67*	*77.1*	*22.9*
5%	*277*	*286*	*66.8*	*33.2*
10%	*376*	*373*	*68.4*	*31.6*
15%	*495*	*472*	*68.4*	*31.6*

This study has clearly demonstrated the usefulness of the decision rule method in assessing the quality of a database. Users can precisely identify normal and abnormal data and compare the quality change across databases. Table 7 summarises the match and unmatched rules between D_1 and D_2 sets for all four databases. Based on the number of defective rules, D_1 and D_2 have the similar data quality problem. The rate for unmatched rules for original database, 5% database, 10% database, and 15% database are 22.9% , 33.2%, 31.6% and 31.6%, respectively, that indicate there are new patterns of quality issues appear in D_2 as the abnormal data increasing.

We also examine the probability of unmatched rules to determine the severity of data quality. In this study, all four databases show no severe quality problems on unmatched rules because the probability of new rules is as small as small 0.01% or maximum as 0.05%.

Table 8. Compare the Rate of Matched rules with P-value

Database	*Matched rules (%)*	*P-value (%)*
Original	*77*	*87.20*
5%	*66.80*	*83.20*
10%	*68.40*	*86.10*
15%	*68.40*	*56.30*

We evaluate the decision rule method against t-test to determine the efficiency of our model. We use t-test to determine a significant deference from D_1 and D_2. To compute t-test, first, we calculate the error rate for each column in D_1 and D_2 for all the following databases: original database, 5% database, 10% database, and 15% database. Then, we calculate t-test to compare the error rate in D_1 with error rate in D_2. The results of p-value for these databases original database, 5% database, 10% database, and 15% database compare with the correspondent decision rules, see Table 8. The results of the comparison in Figure 1, prove that decision rule method has similar attitude to the t-test in these databases original, 5% and 10%. In database 15%, p-value seems to be impacted be the increasing numbers of defective values. Therefore, in database 15%, p-value has different attitude from decision rule.

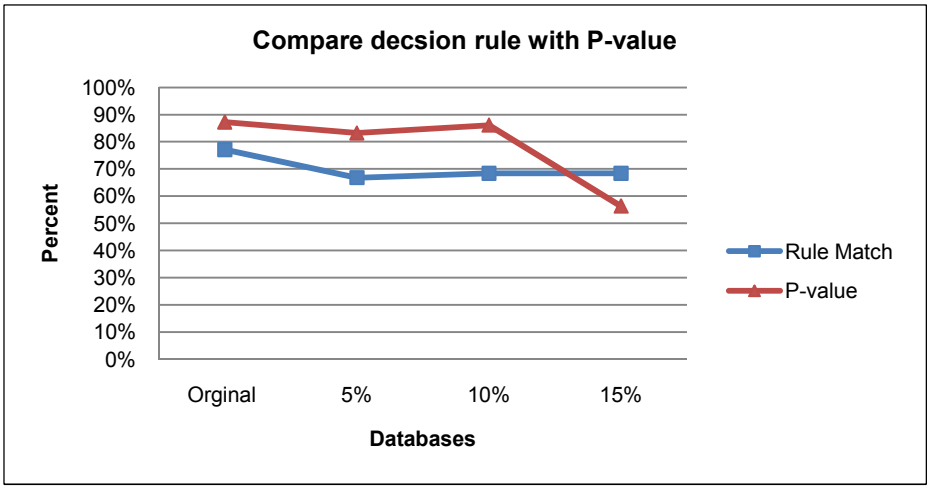

Figure 1. Comparing the Matched rules with p-value

The decision rule can also be extended to measure the randomness degree of errors. Since the number of attributes in this study is 10 attributes, the covering size of decision table is 1024 rules obtain from Eq. (1). Then we compute the covering size of defective rules from Eq. (2) which is 1023 defective rules. From Table 3 for example, the number of defective rules for D_1 and D_2 are 73-1=72 rules and 67-1=66 rules respectively. We measure the randomness degree based on Eq. (3), and the results are listed in Table 9.

Table 9. Randomness Degree Measurement

Database	*D_1(%)*	*D_2(%)*
Original	*7.03*	*6.44*
5%	*26.95*	*27.83*
10%	*36.62*	*36.32*
15%	*48.24*	*45.99*

The results obtained from this study are encouraging and prove the efficiency and effectiveness of the proposed method in assessing the quality of a large size database. Unlike the study proposed in [5], a decision rule method does not have time complexity problem. Additionally, decision rules overcome the limitation existing with error rate method, and can report the reasons of the possible quality problems in databases.

6. CONCLUSION AND FUTURE WORK

Data quality assessment is a critical component for organisational performance as it supports decision makers to make correct decisions that meet organisational needs. Assessing the quality of data also reflects consumers' satisfaction, employees' performance and operational costs. Therefore, it is necessary for management to rely on proper quality assessment.

Most current approaches depend on error rate or accuracy rate to present defective or correct values in a database. However, they do not provide management with the most valuable information with which to conduct quality improvements. Additionally, error rate or accuracy rate is insufficient for quality assessment because errors are randomly and systematically distributed. A recent study [5] extended the error rate method by introducing randomness and probability measurement, yet this method has a problem of time complexity when dealing with a large database. The authors overcame this complexity by segmenting a database into several small segments and computing the average values of each segment. This solution is impractical and inefficient in terms of time and accuracy particularly with a large volume of database.

The decision rule method considers these issues presented on both the method's errors rate and the method presented in paper [5]. The proposed method in this paper provides management with a reliable and efficient data quality assessment and enables decision makers to assess any quality change on organisational information systems early on. By adopting a decision rule method, organisations can examine whether the quality of data improved or defected. Managers and executives can rely on the decision rule to estimate the time and costs required for conducting quality improvement tasks.

A decision rule can also be utilised to compute the randomness degree of errors. Usually, errors are present systematically and randomly in databases. A systematic error has clear and regular patterns, but random errors have no underlying patterns and are randomly spread across columns and rows in various degrees. Therefore, it is necessary from a management point of view to calculate the degree of randomness in order to approximate the time and costs needed to implement quality improvement tasks.

It is common for some rows to have more severe quality problems than others. In this study, a decision rule computes the probability of errors distribution. By presenting the probability of defective data, management can confidently determine which rules have the highest and lowest quality problems.

The decision rule method proposed in this study provides a complete quality assessment, and managers and executives can rely on the decision rule to assess the quality of their information systems. The proposed methods can assist management to determine any quality change across databases, and managers can also depend on the decision rule method to measure the randomness degree and probability of errors in a database. The advantages of this decision rule method will ensure that management can accurately and efficiently assess quality and therefore increase the accuracy of decision making.

For the future, a decision rule method will be empirically assessed in large volume databases to examine its usefulness in providing a complete data quality assessment. Also, a decision rule will be used to present a framework for enhancing the accuracy of a data warehouse.

REFERENCES

[1] Ballou, D. P., & Pazer, H. L. (1985). Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. *Management Science, 31*(2), 150-162.

[2] Ballou, D. P., & Pazer, H. L. (2003). Modeling completeness versus consistency tradeoffs in information decision contexts. *Knowledge and Data Engineering, IEEE Transactions on, 15*(1), 240-243.

[3] Even, A., & Shankaranarayanan, G. (2009). Dual Assessment of Data Quality in Customer Databases. *J. Data and Information Quality, 1*(3), 1-29.

[4] Fisher, C. W., Chengalur-Smith, I., & Ballou, D. P. (2003). The Impact of Experience and Time on the Use of Data Quality Information in Decision Making. [Article]. *Information Systems Research, 14*(2), 170-188.

[5] Fisher, W. C., Lauria, J. M. E., & Matheus, C. C. (2009). An Accuracy Metric: Percentages, Randomness, and Probabilities. *J. Data and Information Quality, 1*(3), 1-21.

[6] Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information quality benchmarks: product and service performance. *Commun. ACM, 45*(4), 184-192.

[7] Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. *Information & Management, 40*(2), 133-146.

[8] Pawlak, Z. (2002). *In pursuit of patterns in data reasoning from data-the rough set way. 3rd International Conference on Rough Sets and Current Trends in Computing*, USA, pp. 1-9.

[9] Pawlak, Z., & Skowron, A. (2007a). Rough sets and Boolean reasoning. *Information Sciences, 177*(1), 41-73.

[10] Pawlak, Z., & Skowron, A. (2007b). Rudiments of rough sets. *Information Sciences, 177*(1), 3-27.

[11] Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM, 45*(4), 211-218.

[12] Redman, T. C. (1996). *Data quality for the information age*: Artech House Boston, MA.

[13] Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM, 41*(2), 79-82.

[14] Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Commun. ACM, 40*(5), 103-110.

[15] Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM 39*(11), 86-95.

[16] Wang, R. W., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. [Article]. *Journal of Management Information Systems, 12*(4), 5-33.

[17] Li, Y., Yang W., and Xu Y. (2006). Multi-tier granule mining for representations of multidimensional association rules, *6th IEEE International Conference on Data Mining*, Hong Kong, pp. 953-958.

[18] Li, Y., and Zhong N. (2003). Interpretations of association rules by granular computing, 3rd *IEEE International Conference on Data Mining*, USA, pp. 593-596.

[19] Li, Y. (2007). Interpretations of Discovered Knowledge in Multidimensional Database, *2007 IEEE International Conference on Granular Computing*, pp. 307-312.

[20] Li, Y. (2008). Data warehousing for association mining, *Encyclopedia of Data Warehousing and Mining, Second Edition, vol 2*, pp. 592-597.