

Comparative Study of Techniques in Reducing Inconsistent Data

Mohd Kamir Yusof and Atiqah Azlan

*Fakulti Informatik, Universiti Sultan Zainal Abidin
Terengganu, Malaysia
mohdkamir@unisza.edu.my*

Abstract

Increasing number of data is occurs because of high demand from organizations to run their daily business operations. Most of organizations have an information system in order to provide quality information to the organizations. In order to provide the quality information, the information system must be able to filter any dirty from data sources. One of the type dirty data is inconsistent data. Inconsistent data is occurs because of data structured from different data sources are different. Four latest techniques to detect and reduce inconsistent data have been identified. These techniques are rough set, logic analysis of inconsistent data, fuzzy multi attributes decision making and functional dependencies of corresponding relation variable. In this paper, these techniques have been studied described with suitable examples. The purpose of studied is to identify advantages, disadvantages and any potential enhancement in reducing inconsistent data from database.

Keywords: *Fuzzy multi attributes decision making, rough set, inconsistent data, quality data*

1. Introduction

Rapid development of technology currently has transformed most data sources from manual to database system. Database can be referred to organized collection of data in digital form. In order to control and maintain the database, there is a software package with computer program known as Database Management System (DBMS) such as Oracle, Informix, DB2, etc. Database must be controlled and maintained as well in order to produce a quality data before generate useful information to users. Dirty data is hot issue in a quality data [9]. Dirty data can be categorized into redundancy data, incomplete data and inconsistent data. However, this paper will address on the issue of inconsistent data. Inconsistent data have been a long standing challenge in database environment. Data sources may conflict with each other at three different level; 1) the schema level 2) data representation level 3) data value level. Data value level inconsistency occurs when there are factual dependencies among the sources in data value that describe the same objects. This problem has risen up the issue where data in the database are no longer reliable and consume more cost and effort. Four latest techniques in reducing inconsistent data have been identified and studied. The purpose of this study is to make an analysis in term of advantages, weaknesses and any potential for enhancement among these existing techniques in reducing inconsistent data. These techniques which are rough set, logic analysis of inconsistent data, fuzzy multi attribute decision making and functional dependencies of corresponding relation variable.

2. Previous Work

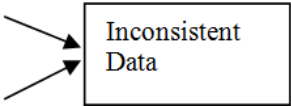
Many researchers have been carried out on the topic of inconsistent data. Each of them has their concept of finding a solution to this problem. A few existing techniques was identified and described in this section.

2.1. Rough Set

The rough set theory was proposed by Pawlak in 1982 as a mathematical tool to deal with vagueness and uncertainty in the classification of objects in a set as mentioned in his research [1]. In this theory provides functionality to generate analyze and optimize set of decision rules obtained from data tables. Data is organized in a table called decision table that consists attribute, notion indiscernible attribute to identify significant attribute as column while as for row, it consists data element, and notion of indiscernible is the discernibility for a subset of objects belonging to domain. The term of indiscernible can be defined as element displayed same information in term of available knowledge. In the research done by Pawlak in 1996, they explained in detail on how rough set theory works. Any union elementary set is labelled as crisp set that can be associate lower and upper approximation. Lower approximation consists of all objects surely belong to set and upper approximation consists of objects that probably belong to set while the boundary is the difference between upper and lower approximation. The concept of lower and upper approximation can be used to deal with inconsistent object that probably or definitely belong to the set [1, 8]. Based on indiscernibility relation concept, redundant features can be identified and eliminated. This is why rough set is suitable for data reduction was proposed by Kusiak in 2001. The following table will serve as running example in this section.

Table 1. Decision Table with Inconsistent Data

U	a1	a2	a3	a4	D
u1	1	1	0	1	1
u2	1	0	1	0	0
u3	1	0	1	1	1
u4	1	0	1	0	1
u5	0	1	1	1	2
u6	0	1	1	1	2



A decision table T is a triplet $T = \{U, A, D\}$ where

$U = \{u_1, u_2, \dots, u_n\}$ – Non empty set of objects (observations, cases or lines)

$A = \{a_1, a_2, \dots, a_n\}$ – Non empty set of attributes

$D = D \subset A$ – Decision attribute

In a rough set table, it allows other value besides the binary values. Note that the table has inconsistent values (u2 and u4) where 2 cases share the same values in all attributes but belong to different decision class.

Given a subset of attributes $B \subseteq A$, $IND(B)$ is called indiscernibility relation of B and is defined as $IND(B) = \{(x, y) \in U \times U : a(x) = a(y), \forall a \in B\}$. In other words $IND(B)$ is a equivalence relation.

Given an arbitrary subset $X \subseteq U$, in Pawlak's rough sets theory (1982) the lower and upper rough approximation, R of X is given by:

$$R_L(X) = \{x \in U : \text{IND}(b) \subseteq X\} \text{ and } R^U(X) = \{x \in U : \text{IND}(b) \cap X \neq \emptyset\}.$$

Following our example, the class $D=1$, $X=\{u1, u3, u4\}$, and the lower and upper approximation are: $R_L(X)=\{u1,u3\}$ and $R^U(x)=\{u1,u3,(u2,u4)\}$.

We can also define the boundary region, $BR(X)=R^U(X)-R_L(X)$, and as a consequence $R_L(X) \subseteq X \subseteq R^U(X)$.

In the example, the decision class is rough since the boundary region is not empty, $BR(X)=\{(u2,u4)\}$. When the lower and upper approximation are equal, $R^U(X) = R_L(X)$, it indicates that there are no inconsistent data and it is called crispy rough set.

In the **reduction by heuristic**, the searching for a core is given by the following procedure: for each iteration, one attribute is removed and the augmenting of inconsistency is checked. If the inconsistency does not grow, the attribute can be removed. When no more attribute can be removed, the remaining ones are indispensable and so the core is found.

By a **discernibility matrix of T**, denoted by M ; we mean an $m \times m$ matrix defined as follow, where $m(i,j)=\emptyset$ denotes that this case does not need to be considered. Following our example, the discernibility matrix M is as follows:

$$M(i,j) = \begin{cases} a \in A : a(ui) \neq (uj) & \text{if } \exists d \in D [d(ui) \neq d(uj)] \\ \emptyset & \text{if } \forall d \in D [d(ui) = d(uj)] \end{cases}$$

Table 2. Matrix M that Consist Inconsistent Data

U	u1	u2	u3	u4	u5,6
u1	-				
u2	a2,a3,a4	-			
u3	s	a4	-		
u4	s	s (inconsistency)	s	-	
u5,u6	A1,a3	a1.a2.a3	a1,a2	a1.a2,a3	-

In table 2, rough set does not exclude or correct the inconsistencies of data but it only allows output discordant decision rules as first and second rules, thus making it more difficult to interpret the result for the end user.

In 2007, new approach was introduced and implemented in order to overcome the limitation in rough set. In this new approach, rough set will check for data consistency based on presented data where it can search for inconsistent data that exist under attribute set in relation of having possible functional dependencies. The main focus in this approach is in relational database where inconsistency usually occurs when redundant data have not been updated unanimously. Functional dependencies that will be used in this process have the property that each right hand side (RHS) of functional dependencies will consist of one attribute. However, for left hand side (LHS) there is no restriction to it [3, 5, 7].

The following table will serve as running example in this section. The relation in the table have two functional dependencies, $\{A, B\} \rightarrow C$, $\{A, B\} \rightarrow D$.

Table 3. Relation Table

Object number	A	B	C	D
1	0	0	1	1
2	0	1	2	1
3	0	2	2	1
4	1	0	1	2
5	1	1	1	2
6	1	1	2	2

The next table will represent will present the data in functional dependency form with values like $A_iB_j \rightarrow C_k$, $A_iB_j \rightarrow D_k$ where A, B, C, and D represent attribute names and i, j, and k represent respective values.

Table 4. Functional Dependencies $A_iB_j \rightarrow C_k$

Object number	FD with values
1	$A_0B_0 \rightarrow C_1$
2	$A_0B_1 \rightarrow C_2$
3	$A_0B_2 \rightarrow C_2$
4	$A_1B_0 \rightarrow C_1$
5,6	$A_1B_1 \rightarrow \{C_1, C_2\}$

Table 5. Functional Dependencies $A_iB_j \rightarrow C_k$

Object Number	FD with values
1	$A_0B_0 \rightarrow D1$
2	$A_0B_1 \rightarrow D2$
3	$A_0B_2 \rightarrow D2$
4	$A_1B_0 \rightarrow D2$
5	$A_1B_1 \rightarrow D2$
6	$A_1B_1 \rightarrow D2$

In table 4 and 5 above, inconsistent data can be detected at object number 5 and 6 for the functional dependency and the degree for the functional dependency is 0.67. While from table 5, no inconsistent data exists in the functional dependency and the degree of dependency is 1.

Based on theory above, inconsistent data can be discovered using this method by applying rough set theory where it measures size of positive region to reflect the dependency between LHS and RHS of functional dependency. Basically, this approach is more practical to use compared to previous approach in rough set. Besides that, this method is more suitable to be implemented in database as it can find some hidden functional dependencies that might be useful for data integrity.

2.2. Logic Analysis of Inconsistent Data (LAID)

Logical analysis of inconsistent data (LAID) is a method developed based on two existing theories, Rough Set Theory and Logical Data Analysis (LAD). The purpose of development this method is to improve the existing theories which are to solve inconsistency created by the process of how sample was developed, that allowed a respondent to belong to more than one class. This new approach caters on the flexibility of rough set and efficiency of LAD [4].

The approach to this new method is almost similar to existing theories, where new test will corresponds to new attribute in the dataset. For each inconsistency detected, a new variable will be added that explains “je ne sais quoi” understood as an indefinable quality. To test this variable, LAD procedures will be used. From here, the link between lower and upper rough approximation and “je ne sais quoi” will be established.

$$jnsq = \begin{cases} 1, & \text{if (boundary region} = 1) \text{ and (class} = 1) \\ 0, & \text{else} \end{cases}$$

The rule for this step is that when two observations are repeated, but belong to different classes, the one variable need to be added. If three or four observations are repeated, then two new “je ne sais quoi” need to be added. Thus, number of unexplained variable is equal to algorithm base 2.

Table 6. The Original Dataset Table

Obs	X1	X2	X3	X4	Class
01	1	0	1	0	0
02	1	1	0	1	1
03	1	0	1	1	1
04	Q	0	1	0	1
05	0	1	1	1	2
06	0	1	1	1	2

Table 7. Dataset after Adding “je se sais quoi”

Obs	X1	X2	X3	X4	jnsq	Class
01	1	0	1	0	0	0
02	1	1	0	1	0	1
03	1	0	1	1	0	1
04	Q	0	1	0	1	1
05,06	0	1	1	1	0	2

Next, the process will apply Disjoint Matrix Procedure where each pair of observations from different classes will be compared. For this matrix, it has managed to overcome the limitation of previous research where the propose disjoint $A[i,j]$ matrix works with an unlimited number of classes. As for the last step, the disjoint matrix that obtained from previous steps will be used as input in the minimum set covering problem, where all constraints must cover at least one by the attribute. This step present an approach named Heuristic approach where for each iteration, a line is chosen to be covered then the best column that covers the column that covers the line and finally solution and remaining are updated. This technique include the inconsistency tolerance and the multiply classes of Rough

Set and the efficiency and attributes cost optimization of LAD. Rough set does not exclude or correct inconsistent data while LAID does not exclude but correct the inconsistencies by adding “je ne sais quoi” variables. The integration of these two approaches is so tight that LAID can be seen as rough extension [4].

2.3. Fuzzy Multi Attribute Decision Making (FMDAM)

Dealing with inconsistent data is one of the challenges in data integration as data that resides at different sources and conflict occur among these different sources. These conflicts may occur at three different level which are schema level, data representation level and data value level. In this technique is focuses on inconsistency at data value level that exists when two or more objects obtained from different data sources are identified similar to each other [2]. These type of inconsistency can only be detected when user request for query. In other to resolve the inconsistent issue, data source quality criteria must be first identified. From there, data model will be developed as it is important for describing and reasoning the contents of data sources.

Definition for data inconsistency:

In query result R , if (1) for object set $OS: \{o_i\}$ ($1 \leq i \leq n$), each object $o_i \in OS$ is obtained from different data sources and refers to the same object RO in the real world and (2) attribute set $\{A_j\}$ ($1 \leq j \leq m$, A is an attribute on local class C and o_i is an object of C) refer to the same attribute of RO and the values of them appear in R and (3) the corresponding attribute in the global classes to all A is GA (4) the corresponding value $o_i.A_j$ of each $A_j \in \{A\}$ are not equal which means GA is not single-valued. Then we say there is a data inconsistency existing in OS of R . And we call attribute set $\{A_j\}$ - inconsistent local attribute set, GA - inconsistent global attribute and OS - polyobject.

Specifically, in polyobject OS of R , for every $oi \in OS$ and $A_j \in \{A\}$ is an inconsistent local attribute set, if $oi.A_j \in \{A_j\}$ where $\{A$ appears in data integration query result, o . Q is collected and recorded for data inconsistency solution. For the global attributes that do not have data inconsistencies, the data source quality criteria vector can be ignored or set to the same value.

The first step to this approach is to obtain fusion matrix where some of the data source quality criteria are quantitative criteria which values are quantitative values such as numbers. Triangular fuzzy number was introduced to represent values of qualitative criteria in order to improve from bipolar scaling method to transform the value of it into triangular fuzzy number.

The next step is to scale the positive and negative criteria value.

$$\text{Equation (1): } Vij = 1 - \left(\frac{qij}{\sqrt{v_i qij + \wedge_i qij}} \right)$$

The third step is to construct fusion decision matrix with an assumption that weight of each data source quality criteria has quantitative value. The fourth step is to compute the distance to the positive ideal solution and negative ideal solution for alternatives.

$$dig = \sqrt{\sum_{j=1}^2 [wj(gj - vij)]^2}$$

$$dib = \sqrt{\sum_{j=1}^2 [wj(vij - bj)]^2}$$

The last step would be to perform fuzzy optimize for data source solution where the degree membership of each candidate data source belonging to the positive ideal solution will be calculated.

$$Set\ vector, = \frac{1}{1 + \left[\frac{dig}{dib}\right]}$$

2.4. Functional Dependencies (FD) of Corresponding Relation Variable

In this technique, association rule finding algorithm was applied. Association rule is based on how often set of items occur together and from this, it will produce information on patterns or regularities that exist in database. Basically, association rule algorithm search exhaustively to find associative patterns therefore many association rules and computing time are required for large target database. This technique is focuses on supplying appropriate minimum support based on target database size. To find inconsistent data in given relation, these are the steps for each user selected functional dependencies in the relation. The first step is to select a functional dependency for data inconsistency check. The second step is to run association rule algorithm for the attributes in the given FD with the parameter of minimum support of 1. The third step is to generate rules for the right hand side of the FD. The fourth step is to find inconsistent data with association rule with confidence less than 100% [3, 5].

3. Summarize of Current Techniques in Reducing Inconsistent Data

Table 8 shows the summary of techniques has been implemented for reducing inconsistent data. The purpose of this comparison is to find out the advantages, disadvantages and potential for improvement in reducing inconsistent data. Some of the researchers have been come out with their idea respectively. Based on these ideas and theories can help other researchers to come with new ideas and theories for reducing inconsistent data.

Table 8. Comparison Techniques in Reducing Inconsistent Data

Techniques	Advantages	Disadvantages	Critics
Rough Set	Able to obtain core and several possible reduction after getting consistent universe -Support many classes and different nominal attribute values	Do not exclude or correct inconsistent data. Allow output discordant decision rules as first and second rules making it difficult for user to	More suitable to be implementation in data reduction and data uncertainty process that involve many classes and attribute values

		interpret the result	
Logic Analysis of Data	Reduce number of attribute in short time	Does not exclude but correct the inconsistencies by adding "jnsq" variables Hardly achieve the goal to identify object based on knowledge of the object.	Involve complex method where for each inconsistent data, a "jnsq" variable will be added
Functional dependencies of corresponding relational variables	Manage to find inconsistent data in short time using functional dependencies between attribute's relation obtained from database	Step to correct inconsistent data are not included. Large dataset need more association rule and computing time Not all database are designed with much consideration about normalization	Bad database design can effect implementation of the technique User need to consider degree of dependency to determine inconsistency
Fuzzy Multi-Attribute Decision Making	Can obtain high average correctness of data of data inconsistency solution	Do not discuss on how to select key that will be used to identify similar objects Data source need to have good quality	Can only be detected while processing user queries Data quality criteria is important element in this technique
Fuzzy Multi-Attribute Decision Making	Has ideal performance compared to other services selection approach	Cannot easily obtained quality of service in open and future pervasive environment	Can only be detected while processing user queries Data quality criteria is important element in this technique

5. Conclusion and Future Work

In conclusion, this paper was described details about current techniques in reducing inconsistent data. The theory and implementation of each technique have been shown and explained. The advantages and disadvantages for each technique also were summarized in table 8. The study is hopefully will give more understanding about techniques can be used in reducing inconsistent data. In future work, the researcher will aim to implement fuzzy multi attributes decision making for reducing inconsistent data from heterogeneous databases. Fuzzy multi attributes was chosen because this technique is better compared to other techniques for reducing inconsistent data.

Acknowledgement

Special thanks to Universiti Sultan Zainal Abidin (UniSZA) my friend Che Mat Ismail for his support and advice.

References

- [1] Pawlak Z, "Rough Set and Data Analysis", Proceedings of the Asian, (1996) Dec. 11-14, pp. 1-6.
- [2] Wang X, Huang LP, Yu XH, Chen JQ, "A Solution for Data Inconsistency in Data Integration", Journal of Information Science and Engineering, vol. 27, (2011), pp. 681-695.
- [3] Sug H, "An Efficient Method of Data Inconsistency Check for Very Large Relations", International Journal of Computer Science and Network, vol. 7, no. 10, (2007).
- [4] Cavigue L, Mendes AB, Funk M, "Logic Analysis of Inconsistent Data (LAID)", (2010).
- [5] Sug H, "A Rough Set Based Data Inconsistency Checking Method for Relational Databases", International Journal of Computer Science and Network, vol. 8, no. 11, (2008) November.
- [6] Tong H, Zhang S, "A Fuzzy Multi Attribute Decision Making Technique Making Algorithm for Web Services Selection Based on QoS", IEEE Asia-Pacific Conference on Services Computing (APSCC '06), (2006).
- [7] Imai S, Lin CW, Watada J, Tzeng GH, "Rough Sets Approach to Human Resource Development of Information Technology Corporations", International Journal of Simulation, vol. 9, no. 2, (2008) May.
- [8] Pawlak Z, "Rough Set", International Journal of Computer and Information Science, vol. 11, no. 5, (1982), pp. 341-356.
- [9] Batini C, Scannapieco M, "Introduction to Data Quality: Data Quality", (2006), pp. 1-18.

Authors



Mohd Kamir Yusof obtained her Master of Computer Science from Faculty of Computer Science and Information System, Universiti Teknologi Malaysia in 2008. Currently, he is a Lecturer at Department of Computer Science, Faculty of Infomatics, Universiti Sultan Zainal Abidin (UniSZA), Kuala Terengganu, Terengganu, Malaysia. His main research areas include information retrieval, database integration and web semantics.



Atiqah Azlan obtained her Degree of Computer Science (Software Development) from Faculty of Informatics, Universiti Sultan Zainal Abidin in 2011. Currently, she is a Master Student at Department Computer Science, Faculty of Infomatics, Universiti Sultan Zainal Abidin, Terengganu, Malaysia.

