A Multi-Dimensional Information Quality Framework to Enhance the Accuracy of Business Intelligence Applications

Mona Nasr, Essam Shaaban, Menna Gabr

Abstract – Data preprocessing is a crucial step through which the data can be cleaned from any quality defects. Quality defects include catching duplicates, filling missing values, removing irrelevant features, catching outliers and other defects. This paper presents a multi-dimensional information quality framework that enhances the accuracy of business intelligence applications by eliminating quality issues in the input data. The results declared that our framework enhances the quality of the data and works effectively.

Index Terms - data quality, quality dimensions, data cleansing, missing values, feature selection, duplication, business intelligence, quality framework.

———————————— u ————————————

1. INTRODUCTION

The use of poor-quality data, having missing and incorrect values, can result in an inaccurate and non-sensible conclusion, making the whole process of data collection and analysis useless for the users. Therefore, in order to deal with the inaccurate and missing values, it is extremely important to have an effective data preprocessing framework [1]. Traditionally, it has been well known that problems related to data quality, such as incomplete, redundant, inconsistent, and noisy data pose a major challenge to data mining and data analysis. In fact, one of the most important steps in data mining is considered to be the data preparation step, which is the process of ensuring the quality of data by changing the original data into a suitable format for the analysis process [2].

· *Mona Mohamed Nasr is Associate Professor, Information Systems Department Faculty of Computers and Information Helwan University E-mail: m.nasr@helwan.edu.eg*
· *Essam Mohamed Shaaban is Assistant Professor, Information Systems Department Faculty of Computers and Information Beni-Suef University E-mail: essam.shaban@fcis.bsu.edu.eg*
· *Menna Ibrahim Gabr is Teaching Assistant, Information System Department Faculty of Business Information Systems Helwan University E-mail: Menna.ibrahim@commerce.helwan.edu.eg*

As information is a vital asset for any business, so the information must be tested against any data quality defects to ensure its effectiveness for use, this assessment is happening in data cleansing step. Data cleansing is a critical step in which data quality assessment is done to remove the quality issues to ensure the high quality of the used data. Data quality refers to how relevant, precise, useful, in context, understandable and timely data is. Data is considered to be of high quality if it satisfies the requirements stated in a particular specification and the specification reflects the implied need of the user [3]. In another way data quality is often defined as 'fitness for use', i.e. an evaluation of to which extent some data serve the purposes of the user [4].

The term data quality is clearly defined and tested through data quality dimensions. Too many data quality dimensions are stated here [5], [6]. For the purpose of this paper we only focus on the Completeness, Relevance, and Duplication dimensions. A simple definition for each quality dimension according to our scope is presented next.

1. Completeness means the extent to which data is not missing and is of sufficient breadth and depth for the task at hand.

2. Relevance means the extent to which data is applicable and relevant for the task at hand.
3. Duplication means a measure of unwanted duplication existing within or across systems for a particular field, record, or data set.

This paper is structured as follows. Section 2 shows the techniques used to solve each quality defect. Section 3 describes the information quality framework. Section 4 demonstrates our experimental results. Finally section 5 provides conclusion and future work.

2. THE SOLVING TECHNIQUES

As we working on Completeness, Relevance, and duplication dimensions, this segment presents the used techniques to solve these quality issues respectively.

2.1 Completeness / Imputation Techniques

Missing data might occur because the value is not relevant to a particular case, could not be recorded when the data was collected, or is ignored by users because of privacy concerns. Missing values lead to the difficulty of extracting useful information from that data set. Missing data are the absence of data items that hide some information that may be important[7]. Knn imputation technique and Multivariate imputation technique using (predictive mean imputation) are used to tackle the missing values problem.

2.1.1 KNN Imputation

In Knn method, missing values are imputed using the most nearest, similar neighbor calculated from the distance function, usually the Euclidean distance. The once the nearest neighbor have been found the replacement value substituted the missing values. The replacement value is calculated based on the type of the data. Euclidean, Manhattan and Minkowski distance are used for numeric data where Hamming distance is used for categorical data.[8]

2.1.2 Multivariate Imputation

Multiple imputation (MI) is a flexible, simulation-based statistical technique for handling missing data. MI as a missing-data technique has two appealing main features: 1) the ability to perform a wide variety of completed-data analyses using existing statistical methods; and 2) separation of the imputation step from the analysis step. MI consists of three steps: 1. Imputation step. M imputations (completed datasets) are generated under some chosen imputation model.

2. Completed-data analysis (estimation) step. The desired analysis is performed separately on each imputation m = 1... M. This is called completed-data analysis and is the primary analysis to be performed once missing data have been imputed. 3. Pooling step. The results obtained from M completed-data analyses are combined into a single multiple-imputation result[9]. The imputation model used with MI is predictive mean imputation.

2.2 Feature Selection Techniques

The main purpose of feature selection is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features. In real world problems, feature selection is a must due to the abundance of noisy, misleading or irrelevant features[10]. Feature selection methods can be broadly divided into filter and wrapper approaches. In Filter approach the attribute selection method is independent of the DM algorithm to be applied to the selected attributes and assess the relevance of features by looking only at the intrinsic properties of the data [11].In wrapper approach the attribute selection method uses the result of the DM algorithm to determine how good a given attribute subset is. The major characteristic of the wrapper approach is that the quality of an attribute subset is directly measured by the performance of the DM algorithm applied to that attribute subset [12]. As filter methods are much faster compared to wrapper methods, and it might fail to find the best subset of features in many occasions but wrapper methods can always provide the best subset of features, therefore we used a hybrid method that combine both Wrapper and Filter technique.

2.3 Duplication Techniques

Heterogeneous data originating from different sources may possibly use different representations of the same real-world entity or concept. The presence of duplicates is a

major problem for maintaining the data quality in large databases[13]. There are two techniques used for matching records with multiple fields namely probabilistic approach and deterministic approach. In this paper we used deterministic approach to catch duplicates.

2.3.1 Deterministic Approach

Deterministic algorithms determine whether record pairs agree or disagree on a given set of identifiers, where agreement on a given identifier is assessed as a discrete—"all-or-nothing"—outcome. It also called "Exact Matching" (requiring an exact match on all identifiers). A record pair is classified as a match if the two records agree, character for character, on all identifiers and the record pair is uniquely identified. A record pair is classified as a nonmatching if the two records disagree on any of the identifiers or if the record pair is not uniquely identified [14].

3. INFORMATION QUALITY FRAMEWORK

As we work in BI environment thus the assessment process is taking place inside the ETL stage, ETL stage has three steps (Extract, Transform and Load). Precisely the quality assessment is happening in Transformation step where we can do data cleansing. Data cleansing or scrubbing is concerned with detecting and removing errors, inconsistencies and other quality problems to enhance the quality of the data. The quality assessment process is passing through four steps as depicted in figure 1: Analysis, Treatment, Assessment and Monitor. In the analysis step we identify the quality problems that we face which in this case are; Missing Values, Relevance and Duplication problems. The identification and recognition of the problems is essential step in knowing what we have in order to know what to do. After the identification of the quality problems we moved to the Treatment step. In treatment step we compare between the available techniques and search for using the best technique to solve quality problems. After that we have to ensure that the used techniques has an effect and solved the problem which make us move to the assessment step. In the assessment step we compare between the dataset before and after using solving techniques and report the results to know the effect of our solution. Observing the percentage of enhancement in data quality reveals by how much we're able to solve the problem and also gives us hints to make re-enhancements. The observation process is done under the monitor step.

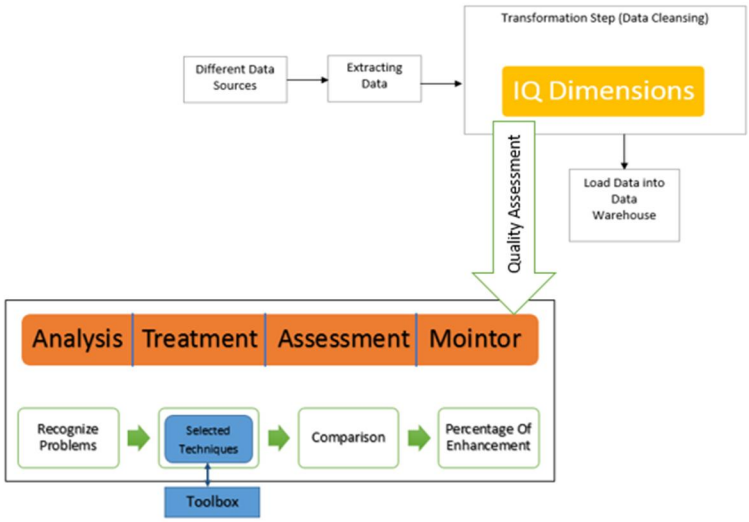

Figure1: The Information Quality Framework.

4. THE EXPERIMENTAL METHODOLOGY

A direct marketing campaigns data set of banking institution is used and downloaded from UCI machine learning repository [15]. It contains 45211 instances and seventeen attributes. Randomly set with 10% sample is selected. Before doing any cleansing operation, classification step using KNN, C5.0, rPart and SVM algorithms is done. The accuracy and error rate for the classifiers are represented in table 1. Then we started to tackle each problem and report the results against the pre-cleansing step's results which is illustrated next.

TABLE 1: ACCURACY AND ERROR RATE FOR EACH CLASSIFIER.

Classification Tool	Accuracy Rate	Error Rate
C5.0 Algorithm	89.9%	10.1%
rPart Algorithm	88.8%	11.2%
Support vector machine(SVM)	88.1 %	11.9%
K-nearest neighbor(KNN)	87.5%	12.5%

4.1 Missing Values Problem

As our dataset doesn't have missing values we created missing values in the dataset by 10% based on the concept of missing at random (MAR). Then we started to impute these missing data using KNN imputation and mean imputation. Two R packages are used to support the KNN imputation (DMwR and VIM packages), each one of them has its own KNN function and results. The Mice and Hmisc packages are used to support multivariate imputation by using predictive mean imputation. We set m=2 and n.impute =2, where m and n.impute identify the desired number of the imputed datasets. After imputing the dataset by the four packages we started to test the performance of each imputation technique by doing classification using four classifiers (KNN, C5.0, rPart, SVM) and reporting the results as shown in table 2. This table represents the average accuracy for each technique. These results depicts that the highest accuracy rate for C5.0 is 90.03 % while KNN reach its highest value at 88.84%. rPart get its highest accuracy rate at 89.98%. For SVM classifier the highest value is 89.17%. Based on the performance results we can conclude that KNN imputation is the best technique to impute missing values. Therefore DMwR package is superior to other packages followed by VIM package.

Table 2: Comparison between Imputation Techniques.

	DMwR (Knn imputation)	VIM (Knn imputation)	Mice (PM imputation)		Hmisc (Mean imputation)	
			1st dataset	2nd dataset	1st dataset	2nd dataset
C5.0	90.03%	89.62%	89.47%	89.49%	89.42%	89.46%
KNN	88.84%	88.78%	87.78%	87.85%	87.68%	87.72%
rPart	89.98%	89.9%	88.98%	89.73%	89.02%	89.22%
SVM	Cost =.1 → 88.98 %	89.17%	88.02%	88.26%	88.12%	88.02%

4.2 Duplication Problem

Regarding duplication problem Two R packages based deterministic matching are used to detect and remove duplicates (data.table and dplyr packages). In our case study we deal with duplicates after data transformation step, after making sure that all the data is in the same format to avoid the case of (ex; 4th and fourth). As the dataset doesn't have duplicates therefore we added some identical rows and other rows with the same values except one value to figure out if they'll be treated as exact matching. Regarding data.table the duplicated function is used to search for duplicates where unique function is used to remove these duplicates. In dplyr package 'which' function is used with duplicated function to return the index of the duplicated records. While distinct function is used to remove duplicates. Both packages remove the identical rows and treat the rows with the same values except one value as non-matched rows. Both packages were able to remove duplicates and return clean data without record copies. Table 3 shows the accuracy rate for each classifier before and after removing duplicates. As represented here the accuracy before removing duplicates is higher than the accuracy after removing duplicates this is because "as the number of instances increase the accuracy increases" but this is not correct values because the data has many records for the same object, so we can't depend on it. But after we removed these duplicates the accuracy rate return to its values as shown in table 1.

Table 3: Accuracy rate before and after removing duplicates.

Classification Tool	Accuracy Rate with duplicates	Accuracy Rate without duplicates
C5.0	90.1%	89.9%
rPart	89%	88.8%
SVM	88.3%	88.1%
KNN	87.7%	87.5%

4.3 Feature Selection Problem

A hybrid feature selection method that combine both filter and wrapper approaches is used in order to gain advantages of both techniques. Filter technique is used first to extract the relevant features and then wrapper technique is applied to select the relevant features from filter subset. The relevant features that resulted from filter techniques are (duration, poutcome, y, pdays, month, previous, age, contact, job, housing, balance, loan and marital).

After applying wrapper technique on this subset we got these features for each classifier. For KNN we got (duration, poutcome, y, month, previous, age, contact, job, balance, loan and marital). Where the relevant subset for C5.0 is (duration, poutcome, y, month, previous, age, contact, job, housing, and marital). The subset for rPart is (duration, poutcome, y, month, previous, age, contact, job, housing, balance, loan and marital). Finally the SVM subset is (duration, poutcome, y, month, previous, age, contact, job, balance, loan and marital). Classification step using the new features is done again to the performance after using hybrid FS method. The results are displayed in table 4.

Table 4: Accuracy Rate After Using Hybrid FS.

Classifiers	Accuracy Rate
SVM	95.4%
C5.0 Algorithm	90.1%
rPart Algorithm	89.1%
K-nearest neighbor(KNN)	88.7%

4.4 Framework Results

After solving each problem separately and concluding best used techniques, we used KNN imputation (using DMwR), hybrid FS method and exact matching respectively on the dataset and record the final results. As shown in table 5 there is an enhancement in the accuracy rate for each classifier compared with the accuracy rate before doing any cleansing steps. Table 6 declared that our framework is working effectively for improving the data quality. As depicted in table 6 there is a huge improvement in the accuracy rate for rpart classifier by 1.7% followed by KNN by 1.5 % then SVM classifier by .8% and eventually comes the C5.0 classifier by 6%. Therefore we can conclude that our framework is working effectively.

Table 5: Accuracy Rate after Solving the Three Quality Issues.

Classifiers	Accuracy Rate
SVM	88.9%
C5.0 Algorithm	90.5%
rPart Algorithm	90.5%
K-nearest neighbor(KNN)	89%

Table 6: comparison between accuracy rate before and after using our framework.

Classifiers	Results before Cleansing	Results after Cleansing
SVM	88.1 %	88.9%
C5.0 Algorithm	89.9%	90.5%
rPart Algorithm	88.8%	90.5%
K-nearest neighbor(KNN)	87.5%	89%

5. CONCLUSION

As presented above the accuracy rate for each classifier is improved after tackling each problem. This means that doing data cleansing step is a focal point to the success of any business especially in a competitive environment like business intelligence environment. The results also declared that our framework is able to increase data quality and make it effective for use. For future work more dimensions can be added to the framework and tested. Also more techniques can be used to tackle these problems and new concepts can be adopted like swarm intelligence and others.

6. References

[1] G. Rahman and Z. Islam, "Missing value imputation using a fuzzy clustering-based EM approach," 2015.

[2] S. Shohdy, Y. Su, and G. Agrawal, "Accelerating Data Mining on Incomplete Datasets by Bitmaps-based Missing Value Imputation," no. June, 2015.

[3] M. Parker, C. Stofberg, R. De Harpe, I. Venter, and G. Wills, "DATA QUALITY: HOW THE FLOW OF DATA INFLUENCES DATA QUALITY IN A SMALL TO MEDIUM MEDICAL PRACTICE," 2006.

[4] A. Haug, F. Zachariassen, and D. Van Liempd, "The costs of poor data quality," vol. 4, no. 2, pp. 168–193, 2013.

[5] F. Sidi, P. Hassany, S. Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, "Data Quality©: A Survey of Data Quality Dimensions," pp. 300–304, 2012.

[6] G. Vasile and O. Mirela, "Data quality in business intelligence applications," *Ann. Univ. Oradea*, pp. 1364–1369, 2008.

[7] Minaksh, D. R. Vohra, and Gimpy, "Missing Value Imputation in Multi Attribute Data Set," vol. 5, no. 4, pp. 5315–5321, 2014.

[8] P. Jönsson and C. Wohlin, "An evaluation of k-nearest neighbour imputation using likert data," *Proc. - Int. Softw. Metrics Symp.*, pp. 108–118, 2004.

[9] StataCorp.Ltd, *Stata Multiple-Imputation Reference Manual*. 2013.

[10] E. Emary, H. M. Zawbaa, C. Grosan, and A. E. Hassenian, "Feature Subset Selection Approach by Gray-Wolf Optimization," pp. 1–13, 2015.

[11] S. Beniwal and J. Arora, "Classification and Feature Selection Techniques in Data Mining," vol. 1, no. 6, pp. 1–6, 2012.

[12] M. L. Raymer, E. D. Goodman, L. A. Kuhn, and A. K. Jain, "Dimensionality Reduction Using Genetic Algorithms," vol. 4, pp. 164–171, 2000.

[13] G. V. Dhivyabharathi and S. Kumaresan, "A survey on duplicate record detection in real world data," *ICACCS 2016 - 3rd Int. Conf. Adv. Comput. Commun. Syst. Bringing to Table, Futur. Technol. from Arround Globe*, pp. 1–5, 2016.

[14] M. B. A. W. R. C. Stacie B. Dusetzina. Seth Tyree, M.S., M.A. Anne-Marie Meyer. Adrian Meyer, M.S. Laura Green, "Linking Data for Health Services," 2014.