Analysis of Aggregation Methods in Incomplete Database Systems

Doña J.M.
Dpt. of Languages and Computer Sciences
University of Málaga.
Spain
jmdona@lcc.uma.es

Quintana O.P.
Dpt. of Computer Sciences
National University of National of the Northeast.
Argentina
oquin@indec.mecon.gov.ar

Valesani M. E.
Dpt. of Computer Sciences
National University of National of the Northeast.
Argentina
mevalesani@exa.unne.edu.ar

Vallejos O. A.
Dpt. of Computer Sciences
National University of National of the Northeast.
Argentina
ovallejos@exa.unne.edu.ar

Abstract

Incomplete data is often a problem present in real datasets and different techniques are used to alleviate this problem. In this paper the performance of the different aggregation methods are analyzed, also OWA operator and traditional imputation technicques are compared.

Keywords: Missing data, Database systems, OWA Operators, Imputation.

1 Introduction

The incomplete databases and missing data is a very common problem in real datasets and different methods to solve this problem have been developed [5, 11].

The experience in databases has demonstrated the dangers of simply removing cases (listwise deletion) from the original data set when incomplete items appear in databases. Deletion can introduce substantial biases in the study, especially when missing data is distributed in a not random way. Missing data values may be frequent in data collection efforts, such as social surveys or scientific experiments, as well as in system data archives. This can be attributed to numerous factors, which include non-response from the sample of the study or malfunction of data collection devices.

The missing data and non-response items can be classified in two groups:

- Records that have all the missing fields.
- Records that have certain fields with missing value.

For the first case the weighting technique is applied, [1, 4] while for the cases in which non-response appears in some fields, the imputation techniques studied in the present work are applied.

The main reason for carrying out aggregation and imputation methods is to reduce non-response bias, which occurs because the distribution of the missing values, assuming it was known, generally differs from the distribution of the observed items. When imputation is used, it is possible to recreate a balanced design so that procedures used for analysing complete data could be applied in many situations [5, 9].

In this sense, the imputation of missing data is an area of statistics which has attracted much attention in the last decades and many different strategies have been developed with the following objetives:

a. To reduce the bias of the estimations.
b. To facilitate the analysis of the database information.
c. To improve the consistency of the results between different types of analysis and to facilitate the process of estimation with auxiliary sources of information.

Undoubtedly, imputation should be applied cautiously and the analysts of the completed data set should be fully warned of the potential dangers created by the imputation. It is very important to reduce the impact of imputed data over the whole of database.

For this reason, it is reasonable to study imputation methods keeping some charac-

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 282–288
Torremolinos (Málaga), June 22–27, 2008

teristics of the variable, i.e.: actual distribution of variable contents, its relationship with the rest of the variables, etc.

This work presents an analysis of the most common aggregation and imputation methods. Also it is determinated the conditions where each method is more efficient in the imputation of data. Finally the OWA operators are studied as imputation operators and compared with the previous methods.

2 Imputation Methods

The solution to the incomplete database problem consists of imputing to fill in missing data with plausible values estimated by means of some method of imputation to produce a complete data set.

During the previous decades there were used procedures of imputation based on the experience, the intuition and the opportunity [11] [12]. Nowadays a great number of methods of imputation are being used and new methods are generated using different statistical skills. Great part of the methods of imputation can be expressed by means of the following equation:

$$y_{vi} = f(y_{nm}) + \varepsilon$$

Where y_{vi} represents the imputed value, y_{nm} is the observations with valid values (not missing), and ε refers to the random residue.

In case of deterministic methods ε is null and it is variable in case of stochastic methods. The deterministic methods (average, medium, etc.) can provide very good results; nevertheless they generate distortions in the distribution of the variable.

Following the characteristics of five usual imputation methods are briefly described using the classification proposed by Laaksonen in [5].

Hot-deck

When the absent information in a record it's found, the hot-deck method replaces it with an existing value in the sample. All the sample units are classified in non-connected groups so that they are as homogeneous as possible in the groups. To every value that is absent, a value of the same group is assigned. The procedure supposes that the lacking information follows the same distribution of those which have values. This supposition incorporates a strong restriction into the model, if this hypothesis is not true the slant will diminish only partly due to the non-response.

The method presents the following characteristics: 1. It allows a post-stratification; 2. It do not present problems at the moment of fitting sets of information; 3. strong supposed are not needed to estimate the individual values of the lacking information; and 4. the distribution of the variable is remained.

Some disadvantages are: 1. It distorts the relation with the rest of the variables; 2. It lacks a mechanism of probability, then it is needed to take subjective decisions that concern the quality of the information; 3. the classes have to be defined on the basis of a limited number of variables, with the purpose of assuring that there will be sufficient complete observations in all the classes; and 4. the possibility of using several times the same answered unit.

Variants of the method:

Hot-deck with random donor: It consists of choosing in a random way one or more donor records for every candidate record. There are different modifications of this method. The simplest case consists in choosing a donor record and then makes the imputation using the candidate record with such information. A sample of donor records can be chosen by means of different types of sampling and make the imputation with the average value obtained with all of them.

Modified Hot-deck It consists in classifying and fitting the potential donors and recipients using a considerable number of variables. Hierarchic bases are used.

Regression Procedures

In this group are included those procedures of imputation that they assign values to the fields to impute, depending on the model:

$$y_{vi} = \alpha + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon$$

Regression is normally used for numeric variables; however for categorical data, logistic regression may be used. A potential disadvantage of predictive regression imputation is the distortion of the shape in the distribution of the variable and the correlation between variables, which are not used in the regression model.

Arithmetic mean (Average)

The arithmetic mean or average of a finite quantity of numbers is equal to the sum of all values divided by the total number of ellements. It is one of the principal sample statisticians. And also it is used as method for the imputation of missing data

Expressed in a intuitive way, the (arithmetical) average is the total quantity of the variable distributed to equal parts among every observation. Also the arithmetic mean can be named as gravity center of a distribution, which is not necessarily the half.

So, given the numbers a_1,a_2, ... , a_n, the arithmetic mean will be:

$$\bar{x} = \frac{\sum_{i=1}^{n} a_i}{n} = \frac{a_1 + ... + a_n}{n}$$

Other statistical averages are: the geometric average, the harmonic average, the quadratic average, the weighted average, etc.

Median (statistic)

In Statistics a median is the value of the variable that leaves the same number of data before and after it. In agreement with this definition the set of minor or equal data that the median will represent the 50 % of data, and those major than the median will represent the other 50 % of the whole sample of information. A medium interval will be the interval containing that piece of information.

Considering x_1, x_2, …, x_n the data of a sample arranged in increasing order and median as: $M_e = x_{(n+1)/2}$, if n is odd then M_e will be the central observation of the values, as soon as these have been arranged in increasing or diminishing order.

$M_e = \frac{x_{n/2} + x_{(n/2)+1}}{2}$, if n is even then M_e will be the arithmetic average of the two central observations.

On having treated with grouped data, if n/2 coincides with the value of an accumulated frequency, the value of the median will coincide with the corresponding abscissa. If it does not coincide with the value of any abcisa, it is calculated by similarity of triangles in the histogram or polygon of accumulated frequencies.

OWA Operators.

An OWA operator of dimension n is an application $F : R^n \to R$, that has an associate weighting vector: $W = [w_1, .., w_n]$ such that $w_i \in [0,1]$, $1 \le i \le n$ and $\sum_{i=1}^{n} w_i = 1$.

Where $F(x_1, ..., x_n) = \sum_{k=1}^{n} w_k x_{jk}$ being x_{jk} the k-th bigger element of the collection x_1, … ,x_n [4]

A fundamental aspect of the OWA operators is the step of the reordering.

An aggregated x_i is not associated with a particular weight w_j, but a weight is associated with a particular ordered position j of the arguments. This arrangement introduces the non-linearity in the process of aggregation (Carlsson and Fullér, 2002).

The generality of this technique is exposed if we show how a great number of operators can be obtained according to the choice of the weights.

Majority Operators

A variant of the OWA operators, are the majority operators, where the aggregation solution is characterized by the elements more representative in the aggregation set without neglecting the opinion of the minority.

As shown in (2) the *MA-OWA* operator is defined as:

$$F_{MA}(a_1, ..., a_n) = \sum_{i=1}^{n} w_i b_i = \sum_{i=1}^{n} f_i(b_1, ..., b_n) b_i$$

where $w_i \in [0,1]$ with $\sum_{i=1}^{n} w_i = 1$. and b_i is the i-th element of a_1,..., a_n arranged in ascending order according to the cardinalities.

The weight of the MA-OWA operator is calculated as follow.

Let δ_i be the cardinality of the element i with δ_i >0, then:

$$w_i = f_i(b_1,\ldots,b_n) = \frac{\gamma_i^{\delta_{\min}}}{\theta_{\delta_{\max}} \cdot \theta_{\delta_{\max-1}} \cdot \ldots \cdot \theta_{\delta_{\min+1}} \cdot \theta_{\delta_{\min}}} + \\ + \frac{\gamma_i^{\delta_{\min+1}}}{\theta_{\delta_{\max}} \cdot \theta_{\delta_{\max-1}} \cdot \ldots \cdot \theta_{\delta_{\min+1}}} + \ldots + \frac{\gamma_i^{\delta_{\max}}}{\theta_{\delta_{\max}}}$$

where

$$\gamma_i^k = \begin{cases} 1 & \text{if } \delta_i \geq k \\ 0 & \text{otherwise} \end{cases}$$

and

$$\theta_i = \begin{cases} (n^o \text{ of item with cardinality} \geq i) + 1 & \text{if } i \neq \delta_{\min} \\ n^o \text{ of item with cardinality} \geq i & \text{otherwise} \end{cases}$$

The MA-OWA is based in majority process where elements with similar values cooperate with other groups of opinion to obtain a representative value of the total group.

3 Analysis

Selected data for the experiment

For the analysis of the previous methods we use the data base of the national agricultural census. This data base has the general characteristic of presenting all constant and discrete numerical variables.

Nine items and 5385 instances are studied. The first item corresponds to a code stratified by size of cattle producer (the only discreet attribute). This item was not considered in the imputation. The item 2, 3 and 9 correspond to numerical information. The rest of the items correspond to types of constant variables.

The method used to delete information in the original database is based in the MCAR (Missing Completely at random).

Design of the Stratification

To assemble the size of the strata bore in mind the quantity of tuplas of the table, and the nature of the data. In the same way, different studies have been done with different size of strata, with grouping of 700 tuplas as the best stratification. The relative errors did not change substantially as the size of the classes was diminishing. The tests of stratification were performed with sizes of 100, 300, 350, 700 and 1000.

The results

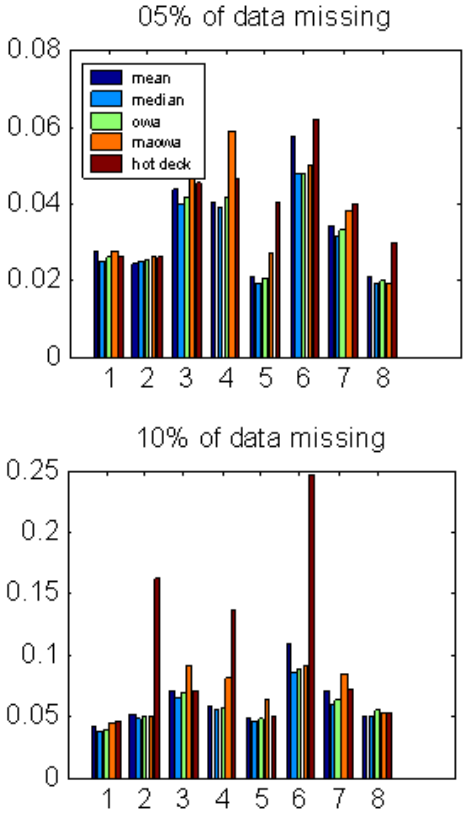
In order to analyze the different imputations regarding their different percentages of missing data, codes and graphs were made. The data are normalized to achieve the analysis of the imputation methods.

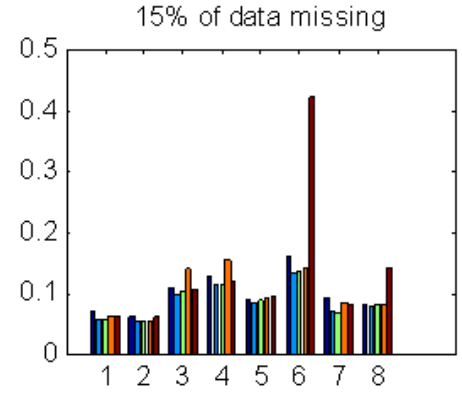
By every execution of the methods, a table of absolute differences between the Real Matrix and Imputed Matrix was generated.

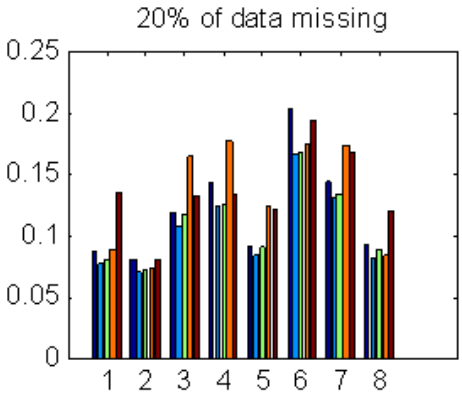
Scripts with absolute errors, relative errors and standard deviation were made. The results were stored in tables and several practices of analytical and demonstrative graphs of every situation were done.

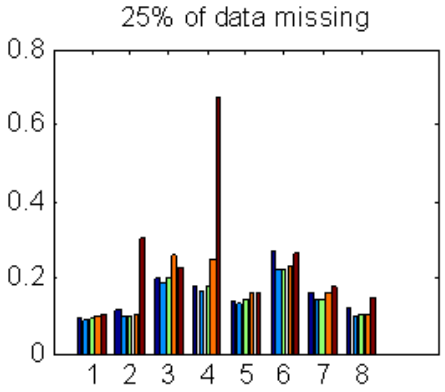
Once every method was codified and tested, we proceeded with the execution of the different imputations on the data base with different percentages of missing data: 5%, 10%, 15%, 20%, 25%, 30%, 40% and 50%.

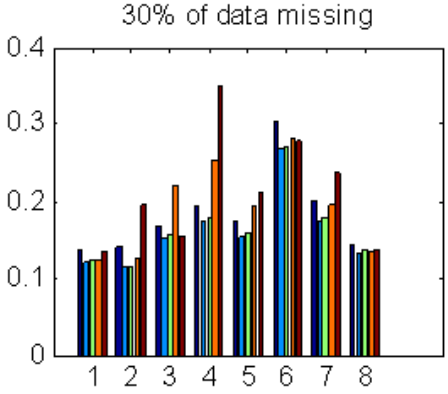
A summary of the results can be seen in the following graphs where the y-axis is the normalized error and the x-axis represents the test for each method:

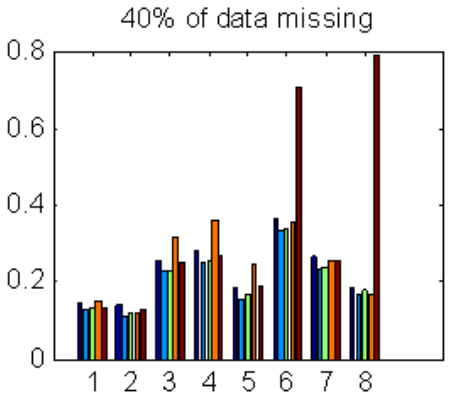


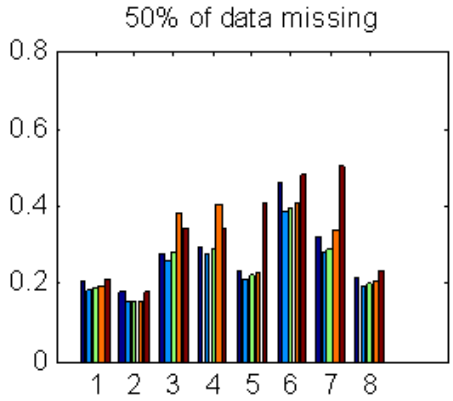












Mean

Being a method of simple imputation, the results were the awaited ones. Being obvious that the best imputation for the item 2 in 4 samples of different percentages of missing data and the worst in the item 6 where it has almost 30 % (2100 reg.) of data with void values (not absent) in all the percentages.

In all the imputations the relative error (RE) was directly proportional to the percentage of missing data.

Another information to observe is that the method behaved relatively well with 50 % of missing data, reaching an average RE of 27 %.

Hot-deck Random

The method got slightly reliable results in all its tests. It provided extreme values in every execution of imputation with a random election of the donor, spoiling the awaited results.

For example, in the case of 15 % of missing data the method produces a maximum RE of the order of 47 %, for 20 % of missing data a maximum RE of 19 % and with only some more 5 % of missing data (25 %) the maximum ER was the order of 67 %, not keeping, in any case,

the increasing proportional relation according to percentages of missing data.

The same relation was produced for the minimal RE. A progressive growth was expected in relation to the percentage of missing data, nevertheless its behavior was unstable in almost all cases.

OWA

Being a method with weighted averages, a better result in the imputations might be expected. Nevertheless it had behaviour similar to the median.

The growth of the ER was proportional to the missing data, yielding to increase from 2% to 4% of difference with the percentage of missing data of the previous order (n-1), unlike the hot-deck that reaches its maximum difference in the order of 12 more points, between 30, 40 and 50 % of absenteeism.

Analyzing imputations of item 6, the OWA behaved satisfactorily, notably improving the percentage of RE with respect to the average and keeping the same behavior of imputation than the method of the median.

The behavior of the OWA demonstrated to be very acceptable, having a difference of error with respct to the median lower than 1 %. It did not have sporadic distortions.

MA-OWA

In [8] is proposed this operator as an alternative for fuzzy imputation. The method, due to the nature of the data, behaved in the desirable way. The method considers the values of the majority and minority. The MA-OWA behaved stabile with data with very big range of variation for some items. Also it produces a RE proportional to the percentage of missing data, reaching a maximum difference of 5% between 30, 40 and 50% of absenteeism.

For the imputations of the item 6, it had a behavior similar to the OWA and to the median, with an error near to 1 % with respect to the other matching methods.

The method increases its RE's relation with respect to its previous imputation (n-1) of missing data in the cases of great absence of information, such the case between 40 and 50 %.

In the cases of numerical imputation with decimal (items 1 and 2), the MA-OWA reaches its best performance.

Median

This is a traditional and simple to apply method, it was one of those that better imputated for these types of data. In the cases of less absenteeism of data its minimal RE was 1,90 %, sharing the podium with MA-OWA.

It reached a maximum RE of 39 % for 50 % of missing data and its minimal RE was 15 %. Bearing in mind the percentage of absenteeism, the method behaved very satisfactorily.

It is necessary to emphasize that for the item 6, the median was the method that best behaved and slightly influenced on the quantity of void information.

The growth of the RE was proportional to the missing data, rising the difference from 2 to 3 % with the percentage of missing data of the previous order (n-1). Also improves slightly the performance reached by the OWA method; and the difference reaches a 4 % only in the tables of major quantity of missing data. Unlike the hot-deck that reaches its maximum difference in the order of 12 points between 30, 40 and 50 % of absenteeism.

If the analysis is focused on the means of the RE it is denoted that the sum of the errors of imputation relative to the real aggregation of every item is low in all the imputation methods.

When it is considered in a percentage of lack for the extremes test, in general lines, the method that had minor relative error was the Median.

The absolute error for all the used methods behaved in a very similar way in all the items, except for the HotDeck Random method that increases the absolute error in an important proportion in the items 5 and 8 with 50 % and 30 % respectively.

4 Conclusions

This paper discusses a range of imputation methods to compensate for missing data and item-nonresponse in data base systems, and illustrates advantages and disadvantages of the methods in a real system.

The produced results had shown the importance to consider the type of analysis and the type of point estimator of interest when applying

imputation. In particular, it should be distinguished if the goal is to produce unbiased and efficient estimates of means, totals, proportions and official aggregated statistics or a complete micro-data file that can be used for a variety of different analyses and by different users. Also, the analysis contributes to present OWA operators as an alternative to incorporate semantics and concept like majority in imputation methods.

The combination of the characteristics of the OWA and majority operators with the imputation methods can improve the traditional systems for the data missing problem.

Acknowledgements

This work is supported by the Project TIN2006-14285. Ministery of Education and Sciences. Spain

References

1. Ezzati-Rice TM, Khare M, Rubin D, Little R, Schafer J. A comparison of imputation techniques in the Third National Health and Nutrition Examination Survey. Proceedings of the Section on Survey Research Methods. American Statistical Association. 1994.
2. Gómez J., Palarea J. Métodos de inferencia estadística con datos faltantes. Estudio de simulación sobre los efectos en las estimaciones. ESTADÍSTICA. 48, 162: 241 – 270. 2006.
3. Grajalesa L. F., López L. A. Data Imputation in Switchback Designs Using a Mixed Model with Correlated Errors. Revista Colombiana de Estadística 29 (2): 221-238. 2006.
4. Jinn J.-H., Sedransk J. Effect on Secondary Data Analysis of Common Imputation Methods Sociological Methodology, 19. 213-241. (1989).
5. S. Laaksonen. Regression-based nearest neighbour hot decking. Computational Statistics, 15 (1):65-71, 2000.
6. Little, R., Rubin, D. B. Statistical Analysis with Missing Data, 2º ed. Jonh Wiley & Sons. 2002.
7. Lorga Da Silva A., Saporta G., Bacelar-Nicolau H.. Missing Data and Imputation Methods in Partition of Variables. Classification, Clustering and Data Mining Applications, D.Banks et al. editors,, pp. 631-637, Springer. 2004
8. Peláez J.I., Doña J.M. Majority additive-ordered weighting averaging: a new neat ordered weighting averaging operators based on the majority process. International Journal of Intelligent Systems. 18(4):469- 481. 2003.
9. Peláez, J.I., Doña, J.M., La Red, D. Fuzzy Imputation Method for Database Systems. Handbook of Research on Fuzzy Information Processing in Databases. Hershey, PA, USA. 2008
10. Schafer, J.L., Graham, J.W. Missing Data: Our View of the State of the Art, Psychological Methods,7, 2, 147-177. 2002.
11. Yager R. Families of OWA operators. Fuzzy Sets and Systems. 59:125-148. 1993.
12. Yuan Y. C. Multiple Imputation for Missing Data: Concepts and New Development. SUGI Proceedings, 2000.