Maximising the Value of Missing Data

Atai Winkler

Received: 13 December 2013 Accepted: 1 January 2014 Published: 15 January 2014

Abstract

The subject of missing values in databases and how to handle them has received very little attention in the statistics and data mining literature1, 2, 3 and even less, if any at all, in the marketing literature. The usual attitude of practitioners is 'we'll just have to ignore records with missing values'. On the other hand, a few very advanced theoretical solutions have been developed, some of which have been applied, particularly to clinical trials data. These solutions can only be applied to small databases, not to the very large databases held by many companies on their customers. This paper describes a new method for imputing missing values in such very large databases. Two particular features of the method are that it can handle all combinations of variable type (continuous, ordinal and categorical) and that all the missing values in the database are imputed in one run of the software. It is based on the k-nearest neighbours method, a well known method in data mining. The paper concludes by presenting the results of a study of this method when used to impute the missing values in a real set of data.This paper is only concerned with 'missing' data, i.e. data that are not known but which have real values. It does not address the problem of 'empty' data, i.e. data that are not known but which cannot have real values.

Index terms— missing data; imputation; gaps; holes; data mining; empty data

1 Introduction

arketers are always striving to develop effective marketing campaigns by maximising the benefits they can achieve from the data they hold in their databases. The results of different campaigns can then be assessed by comparing the return on investment of the campaigns. Whatever the nature and aims of the campaigns, they always start with some form of analysis to gain customer insight, and the results of this analysis lead to segmentation and targeting of potential customers.

As companies move through the cycle of customer acquisition to customer retention, the ability to analyse the data they hold on their customers becomes increasingly important. However, this task is often hampered by the fact that the vast majority of databases have missing information, sometimes called gaps or holes. This may be for historical reasons where the emphasis was on customer acquisition or it may be due to changes in the needs of the organisation over time. Whatever the cause, missing data are a very common and serious problem -it is not uncommon for collected lifestyle data to have 30-40% of their values missing. This is a real problem Author : e-mail: atai.winkler@win-tech.co.uk when the data are to be analysed -if the data are not there, they cannot be analysed either at record level or for the database overall. These missing values mean that the database is not as large or as rich in information as may be assumed from its overall size (number of fields and number of records). Thus, the effect of the missing data is to limit the amount and quality of new information that can be learnt from the database.

With respect to customer retention programmes and CRM, missing customer data will have a serious adverse impact on the outcome of the campaigns because they will be based on incomplete data and therefore weak analysis. The consequence of this is that the segmentation and targeting will be less accurate than they would be if the database were fully populated. This problem of missing data has also heavily affected the lifestyle data sellers who are trying to present a more complete picture of the UK adult population. Since less data are acquired now than formerly directly from individuals about their demographics as consumers, there is a greater reliance on modelled data. Unfortunately, these modelled data are often obtained from incomplete data. This just compounds the problem of missing data because any bias in the available data manifests itself, and is probably increased, in the modelled data. Thus, a vicious circle of partial data being used to obtain more partial data is started, and after this process has been repeated many times it is likely that the final data bear little resemblance to the true UK population.

Since the problem of missing values in both customer and lifestyle databases is widespread and getting worse, methods that give more accurate estimates of the missing values compared to those obtained using currently available methods have an important and significant contribution to play in improving marketing effectiveness. The desired result of all these methods is a fully populated database with all the missing values replaced by estimates of their true values -a process known as (missing value) imputation. The imputed values should be 'good' and plausible estimates of the true values so that the statistical properties of the fully and partially populated databases are as similar as possible.

2 II.

3 Missing Data and Empty Data

It may be thought that all missing data can and should be imputed. This is not always the case because some data may be missing because they cannot have real values -such values must remain missing. Thus, it is important to understand the difference between 'missing' data and 'empty' data. A value that is not known but which has a real value is a 'missing' value. A value that is not known but which cannot have a real value is an 'empty' value. Therefore, missing data can be imputed but empty data cannot be imputed.

4 III.

5 Imputation Methods

A number of imputation methods are available, and some of the factors that help determine a suitable method are:

? Is the method only suitable for small databases or can it be used on small and large databases? Case deletion avoids rather than solves the problem of missing values because it ignores all the incomplete records. Very often it is the default method of handing missing data. Although very easy to implement, two immediate and severe disadvantages of the method are firstly that a very large proportion of the records may be ignored and secondly that the remaining records may not be representative of the population. The commercial and financial implications of this bias in the data that are actually analysed are easy to imagine.

6 c) Mean Substitution and Cold Deck Imputation

Mean substitution and cold deck imputation are two frequently used imputation methods. Mean substitution involves replacing all the missing values in each field by the field's mean or mode as appropriate, and in cold deck imputation the missing values are replaced by external constants, one for each field. These methods are easy and quick to implement but being global methods they are very unlikely to maintain the statistical properties of the database. In the case of mean substitution, the mean (mode) values of the fields in the partially and fully populated databases are, by definition, the same, but the variation of each field in the fully populated database is much smaller than the corresponding variation in the partially populated database. The result is that the records are not as clearly differentiated as they should be and so it is harder to understand how people's individual characteristics determine their actions and behaviour from a database which has been fully populated in either of these ways.

Another major problem with mean substitution and cold deck imputation is that unrealistic or even impossible values can be easily imputed. This is because the value imputed in any one field is the mean of the known values in that field. Therefore, if a database contains people across a wide range of age, income and lifestyle attributes and the data can be segmented into a finite number of homogeneous clusters with high inter-cluster heterogeneity, the mean value of any field across all clusters does not have meaning or significance for any single cluster or all the clusters. Therefore, using values imputed in this way as the basis for marketing campaigns and other commercial activities may not yield the desired outcomes because the targeting and segmentation are based on poor quality dat. A slightly more advanced method of imputation is hot deck imputation. This is similar to cold deck imputation except that the missing values are replaced by values from 'similar' records in the database. These similar records are obtained by clustering the complete records and then assigning a cluster to each incomplete record. The missing values in each incomplete record are replaced by values calculated from its associated cluster. Like mean substitution and cold deck imputation, hot deck imputation is a global method.

7 e) Regression

In regression imputation the missing values are replaced by values calculated from a regression equation, for exampley = a + bx1 + cx2 (1) y is the variable to be imputed, and x1 and x2 are other variables (a , b and c are known constants).

Implicit in using (**??**) is that the values of the variables on the right hand side of it (x1 and x2) in records whose values of y are to be imputed are known. This problem can be overcome by developing the models only from complete records -but this raises a fundamental problem, namely what happens if the complete records are either a small proportion of the database or they are a distinct group in the database rather than being a fair reflection of the database as a whole? On the other hand regression imputation is a local method because the missing values in each record are calculated from the data in that record -a significant advantage. Notwithstanding this advantage, regression imputation has a number of practical and theoretical problems, including:

? since a regression equation must be developed for each variable with missing values, regression imputation is very time consuming, especially in large databases and in databases many of whose fields have missing values; ? working out the equations may be difficult, not least because the correlations between the variables may be weak; ? different relationships may exist for different homogeneous groups in the database and so trying to find one relationship across all groups will yield an unsatisfactory compromise that is not an accurate portrayal of the relationship in any one group -the single equation will predict values that do not reflect any individual's unique characteristics, and so the same problems as those associated with mean substitution, namely reduction in the heterogeneity of the database, may arise;

? the relationships between the variables are artificially and falsely inflated because the missing values are estimated by substituting into the regression equations.

8 f) EM Algorithm and Structural Models

These methods are very advanced and demanding in terms of the time and expertise required. They are not amenable for use on large databases. k-nearest neighbours is a data mining method used in estimation and classification problems. Unlike many other methods used in statistical data analysis and modelling, it does not require a model to be developed for each field. Rather, it is based on the simple concept that the (statistical) similarity between two records is calculated from the multivariate distance between them. If two records are similar, i.e. their corresponding fields have similar values, they will be close to one another and so the distance between them will be 'small' when their common known data are plotted. These records are more similar to one another than are other records with larger distances between them. This geometric way in which the most similar records are found explains why the method is called k-nearest neighbours. Thus, the method involves mapping all the data into multi-dimensional space and then calculating the distances between all pairs of records (each dimension is a variable).

The method works by firstly finding a pool of donors, i.e. complete records, for each recipient, i.e. incomplete record. It then uses the values in the donors of each field that is missing in the recipient to impute the missing data in the recipient. There are three stages to the method. For each incomplete record: 1. Search the entire database for similar complete records using the values in the selected fields in the incomplete record. 2. Rank the complete records by distance to the incomplete record. 3. Use the specified number of complete records in the ranked set to impute the missing values in the incomplete record. By using all the known data in each recipient to search for its most similar donors and then using these donors to impute the recipient's missing values, the statistical properties of the partially populated database are maintained. This process of searching for similar donors and then using them to impute the missing values is repeated for each recipient. This recipient-by-recipient approach means that each recipient has its own donors, i.e. Nearest Neighbours (NNs), from which its missing values are imputed. It is this feature of k-nearest neighbours that helps maintain the heterogeneity of the database. This very localised approach to imputation is in marked contrast to global methods where the imputation is based on groups of recipients and each group has the same donors and therefore the same imputed values. Thus, the variation of the variables in a database which has been fully populated using a global method is lower than it is in a database which has been fully populated using k-nearest neighbours.

Furthermore, since each recipient record is treated individually, the method obtains the most accurate imputed values for each recipient record rather than attempting to obtain the most accurate average imputed values across a group of records.

The main reason for the limited use up to now of the k-nearest neighbours method with very large databases is that the number of distances that have to be stored and then ranked made it impractical to use on such databases. This problem has now been overcome so that it does not store the distance from the incomplete record being processed to each complete record, rank all the distances and then select the specified number of NNs. This means that only a fraction of the number of complete records in the database are stored at any one time.

A good example of the type of data that can be imputed using k-nearest neighbours is lifestyle data.

However, variables such as ownership of pets, type of credit card owned and participation in hobbies, for example stamp collecting, should not be imputed because they are generally independent of other variables and do not define people.

To The table shows that the three NNs for record 3 are records 4, 7 and 2, with record 4 being the most similar and record 2 the least similar of the three NNs. Since there are six complete records in the database, there can be up to six NNs.

If only one donor (the NN) is to be used, the missing values in records 3, 5, 8 and 10 are copied directly from the corresponding fields in records 4, 1, 2 and 9 respectively. If two NNs are to be used, the missing values in records 3, 5, 8 and 10 are calculated from the corresponding fields in records 4 and 7, 1 and 2, 2 and 7, and 9 and 1 respectively. If three NNs are to be used, the missing values in record 3 are calculated from the corresponding fields in records 4, 7 and 2, record 5 from records 1, 2 and 4, record 8 from records 2, 7 and 4, and record 10 from records 9, 1 and 2.

The fully populated database shown in Table **??** was obtained by using the three NNs shown in Table 2.

As with many methods of data analysis, the trade-off between run-time and accuracy must be considered -a big increase in run-time is not always accompanied by a significant improvement in accuracy.

For k-nearest neighbours an obvious question is how the number of donors affects the accuracy of the imputed values. It is reasonable to assume that as the number of donors increases, the 'better' will be the imputed values. However, increasing the number of donors has two adverse effects: firstly, as the distance between the recipient and the donors increases, the donors become more dissimilar from the recipient; and secondly, the run-time increases. This and other questions are discussed later in this paper.

9 Example of Using k-Nearest Neighbours For Imputation:

The use of the k-nearest neighbours method for imputation was tested on a database of 20,000 records, of which 13,303 (66.5%) were fully populated and the remaining 6,697 (33.5%) had at least one missing value. The actual values of these missing values were known so that after the imputation the imputed values could be compared with the actual values. Table **??** Table **??** The table had 8 fields, as described in Table **??**.

Four runs were carried out. The settings of the runs are shown in Table **??**.

10 h) Imputation Results

The results of the imputation for the text variables are shown in Table **??**.

Volume Income was the only ordinal field. It was divided into 10 levels, and its null count was 1,255. The results of the four runs are shown in Table **??** as the percentage of entries in cells off the Leading Diagonal (LD) in the cross classification matrix (this is just a crosstab of actual values against imputed values). Table 10 IV.

11 Discussion of Results

Table **??** shows that the percentage of correctly imputed text variables is not very sensitive to the sampling percentage or to the number of NNs. This is because the records are randomly distributed in the database rather than there being a number of distinct and homogeneous groups in the database each of which is concentrated in adjoining records, and the variation of the fields in the NNs hardly increases as the number of NNs increases.

It is expected that as the number of NNs increases the variation of the fields in the NNs would also increase because the (complete) records in the search domains are ranked by similarity to the incomplete record. This means that the record most recently added to the search domain is always most similar to the previously added record and least similar to the first record in the search domain. If the variation of the fields in the NNs were much greater, the success of the imputation would be much more sensitive to the number of NNs. In general, the rate at which the variation of the fields in the NNs increases depends on the data and the number of NNs.

Another interesting result in Table **??** is the variation of the percentage of values correctly imputed across the fields; for example in run 1 it is between 37.43 and 90.62. Now, it is reasonable to assume that as the number of categories of a text field increases, the percentage of values correctly imputed would decrease (all other things being equal). However, there is another more important factor at play, and that is the standard deviation of the relative frequencies of the categories. Table 11 shows this standard deviation, the rank of the standard deviations, the rank of the percentage of values correctly imputed for any of the four runs in Table **??** and the rank of the number of categories for all the text fields (1 is the smallest rank and 6 is the largest rank). There is an almost perfect 1 to 1 correspondence between the third and fourth columns in Table 11.

12 Run

The only discrepancy occurs with the ranks of h_shrs and h_prop -their standard deviations are more similar to one another than are other pairs of standard deviations. This suggests that the variation in the relative frequencies of the categories of the fields is a significant factor in determining the percentage of values correctly imputed. However, the fourth and fifth columns in Table 11 appear to follow a weak inverse relationship -the larger the number of categories a text field has, the smaller is its percentage of values correctly imputed likely to be.

The results in 12.

Table **??** shows that for all the runs about half the values were imputed correctly or within one level either side of the LD, and that the percentages fall off very quickly as the cells move away from the LD.

Comparing Tables 6 and 10, it is immediately apparent that the run-time is strongly influenced by the number of complete records from which the imputed values are calculated and not affected at all by the number of NNs. This is as expected because the number of distances calculated is given by the product of the number of incomplete records and the number of complete records used. The number of NNs does not determine the run-time because the software used in this study has a very powerful sorting algorithm which sorts the distances as they are processed. This means that the number of distances actually stored as each incomplete record is processed is kept to a minimum and is very close to or equal to the number of NNs specified at input. The alternative solution to this sorting problem is to store the distance for each

13 Position

No records have been processed rank all the distances complete record and then when all the complete from smallest to largest. This approach has two big computational problems: firstly, the larger the database the more distances have to be stored; and secondly the time required to sort these distances is not insignificant, and indeed may be greater than the time required to calculate the distances and impute the missing values.

V.

14 Conclusion

This paper has presented the results of a study on how a powerful data mining technique, k-nearest neighbours, has been enhanced and then applied to the ever-present problem of how to impute missing values in large databases. The enhancements make the method amenable for use on very large commercial databases. The results of a study presented in this paper show that its overall accuracy is very high. The advantage of having more accurate imputed data is that campaigns can be better targeted.

In turn, this will generate higher response rates and a greater return on investment. [1] [2]

2

show how k-nearest neighbours works,
consider the data in Table 1. The data come from a survey and the fields are:
mar_stat: marital status: D (divorced); M (married); S (single)
res_stat: residential status: P (owner-occupier); T (rent alone); Z (multiple rent)
age: age (months)
bank: time with bank (months)
cheq_card: own a cheque guarantee card: N (no); Y (yes)
add: time at current address (months)
emp: time with current employer (months)
occup: occupation code

Figure 1: Table 2 .

1

Figure 2: Table 1

2

Figure 3: Table 2

	Field	Description	
	p_hhld	head of household	
	h_comp	household composition	
	h_shrs	value of shares held	
Year 2014	h_prop h_inc h_res h_ten	type of property household income residence type household tenu	
46	age		age
	Run No.	Sampling Percentage	No. of Complete No. of Nearest Records Sampled Neighbours
	1	10	1,300
	2	10	1,300
	3	5	665
	4	5	665

D D D D) c
(

Field	Null	No. of	% Correctly	% Correctly
Count p_hhld	991	2	90.62	90.41
h_comp	1,090	12	37.43	35.23
h_shrs	1,090	3	76.97	77.34
h_prop	1,189	5	83.52	83.35
h_res	1,255	5	44.94	45.10
h_ten	1,189	3	68.04	66.69

Figure 4: Categrs. Imp. (Run 1) Imp. (Run 2) Imp. (Run 3) Imp. (Run 4)

5678

Run No.	ME	MAE
1	0.64	12.69
2	1.15	12.82
3	1.07	12.90
4	0.71	14.21

Figure 5: Table 5 Table 6 Table 7 Table 8

11

	% Corr % 1 off % 2 off % 3 off % 4 off % 5 off % 6 off % 7 off % ?8 off								
No.	Imputed	LD	LD	LD	LD	LD	LD	LD	LD
1	17.53	32.91	25.02	14.02	6.93	2.63	0.72	0.24	
2	17.05	31.39	24.86	16.10	7.09	2.71	0.80		
3	15.62	31.08	23.59	16.02	8.84	3.82	0.80	0.23	
4	17.13	29.72	20.88	15.30	9.24	4.46	2.07	0.88	0.32
Run No.	Run-Time (Mins)								
1		47							
2		48							
3		24							
4		24							

Figure 6: Table 11 Field St. Dev Rank of Rank of % Rank of No. St. Dev Corr. Imp. Categories

Year 2014
48
Volume XIV Issue III Version I
D D D D) c
(
Global Journal of Computer Science and Technology

Figure 7:

12

	. of Values
LD	10
1 off LD	18
2 off LD	16
3 off LD	14
4 off LD	12
5 off LD	10
6 off LD	8
7 off LD	6
8 off LD	4
9 off LD	2

Figure 8: Table 12

.1 Global Journals Inc. (US) Guidelines Handbook 2014

www.GlobalJournals.org

[Batista and Monard ()] 'An Analysis of Four Missing Data Treatment Methods For Supervised Learning''. G E A P A Batista , M C Monard . *Applied Artificial Intelligence* 2003. 17 p. .

[Batista and Monard ()] *K-Nearest Neighbour as Imputation Method: Experimental Results*, G E A P A Batista , M C Monard . 2002. University of Sao Paulo (Technical report)

[Manly ()] *Multivariate Statistical Methods, A Primer*, B Manly . 1994. Chapman and Hall. p. .