A Comparison of Imputation Techniques using Network Traffic Data

Fidan Kaya Gülağız
Computer Engineering
Kocaeli University
İzmit, 41380, Turkey

Onur Gök
Computer Engineering
Kocaeli University
İzmit, 41380, Turkey

Adnan Kavak
Computer Engineering
Kocaeli University
İzmit, 41380, Turkey

ABSTRACT

Creation of data sets to be used for studies in many different fields of research is really important process. However these data sets suffer from the problem of missing values. There are many different ways of handling missing values. Deletion methods and single imputation methods are the most common ones of these methods. However, this methods lead to high errors in data sets with high loss rates. Data sets used for the analysis of network traffic are also commonly encounters with the missing values. In this study, data produced in different sizes and different missing value rates for the analysis of network traffic in distributed systems. Then, different data imputation methods are compared for dealing with missing values in these datasets. Experimental results showed that Expectation Maximization Method is more applicable and performs better at relatively high missing data rates and k Nearest Neighbors Method performs better at low missing rates.

Keywords

Least Square Estimation (LSE), Expectation Maximization (EM), k Nearest Neighbors (k-NN), Traffic Data, Missing Value Imputation.

1. INTRODUCTION

Data mining is a technique used for extracting hidden patterns of within the data or obtaining useful information from data. One of the most important topics in data mining is the accuracy of the recommendation or method that is used. However, losses in the data sets that is worked on will affect the accuracy value of the data. There are many different methods used for obtaining the missing value. Each of these methods has various advantages and disadvantages. Studies have been undertaken for many years in order to show the superiority of these methods in comparison to each other.

Giraldo et al. have [1] used data sets that belongs to eighteen different fields in order to compare the imputation techniques. As a result of the performed studies, it can be said that the imputation methods do not have clear advantages in comparison to each other if compared alone, however it has been demonstrated that the best result is obtained when the methods are used together with the nearest neighbor method. Bhekisipho et al. [2] studied the relationship of the data imputation methods together with the distribution of the missing values in the data. In conclusion of the study, they demonstrated that the method of deletion missing data makes the data low-quality that the most successful method is the multiple imputation method and that methods such as decision trees are highly affected by the distribution of the losses in the data set. Ten different imputation methods were compared by Gang and Tongmin [3] in order to obtain the missing values present in the daily traffic data. As a result of the performed study, the most successful methods which can be used for the losses in the traffic data were determined as LSI_gene, LSI_array, LSI_combined, LSI_adaptive, EM_gene and Local Least Square Imputation methods. Chih - Feng et al. [4] in the other hand compared the deletion and single imputation methods. The study pointed out that the EM method displays very good performance in case there is missing data which has one property and that the k-NN method displays much better performance if the losses have more than one properties. Yuebiao et al.[5] have categorized the imputation methods in three different categories; prediction methods, interpolation methods and statistical learning methods. The performed study proved that the methods display varying result of different loss ratios with different data sets.

If the comparisons of the previous studies are investigated, it can be observed that there has been a focus on more known data sets. In this study, a comparison of the LSE, EM and k-NN methods has been made for the obtainment of missing values occurring in a data set which has been created in relation to the packages sent to distributed network architecture.

The rest of the paper is organized as follows. Patterns of missing data sets are presented in Section 2. Section 3 describes the data imputation approaches, which will be used in this study. Section 4 presents the network traffic dataset, Section 5 presents the testing results and finally, the conclusions are summarized in Section 6.

2. PATTERNS OF MISSING DATA

The missing data is classified in three different ways according to their occurrence in the data set. These are; missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) [6].

MCAR mechanism is the condition where the randomness is highest. Data loss situations are independent of other observed or missing data. An example for this mechanism is loss of several of the obtained data due to various reasons. The MAR mechanism is a different limited variation of MCAR. MAR mechanism is not dependent on the missing data. Intentional skipping of the asked questions in a survey study by the participants or intentional wrong answers can be given as examples for the MAR mechanism. However, the NMAR mechanism depends on the missing data and the formation of missing data is based on the non-measurements of the units or occurrences by the researcher. This situation may seriously affect the result of data analysis. The most basic example for this situation is a wrongly asked question in a survey which is answered wrongly [7].

MCAR and MAR mechanisms are suitable types for the application of data imputation methods. In the NMAR method, the registered data has wrong values so, it is very difficult to obtain this missing data via any method. A model which is appropriate for the missing data has to be established at this point.

The data presented in the data sets which were produced within the context of this study are MCAR type. Analyses were performed with data which have different loss ratios while evaluating the data sets. The following formula [6] was used during the calculation of the missing data ratios present in the data sets.

$$\mathrm{p} = \frac{\text{the number of the missing data points}}{\text{the number of the total data points}} \tag{1}$$

3. DATA IMPUTATION APPROACHES

There are too many different methods in data mining to complete missing values. Each of these methods complements the data in different ways. Basically, these methods can be listed as ignoring the records with missing values, replacing them with a global constant, filling these missing values manually based on your domain knowledge, replacing them with the variable mean (if numerical) or the most frequent value (if categorical), using modelling techniques such as Nearest Neighbors, Bayesian Rules, Decision Tree or EM algorithms [8].

In this section detailed explanation of missing data imputation methods that are used in this paper are given.

3.1 Least Square Imputation Method

All data can be summoned an analyzed in a table during the finding process of missing data within a data set and a function that models the data may be searched as well. Most of the time, it is not possible to find a function that is completely compatible with the data sets, however the most suitable function may be determined. The process of finding the most suitable function of a data set is called as regression analysis. One of the most frequently used methods during the regression analysis is the least squares method.

There are two ways of obtaining missing data with the least squares method [9]. These are; non-missing data model approximation and complete data model approximation. During the model establishment of the first approach, only data with no missing properties is used, whereas in the second approach data with missing properties is used as well. Here, a model update is being performed via an iterative approach where the property values obtained in each iteration are added to the data. The non-missing data model approximation was used during the establishment of the model in the study.

Let us assume that m records are presented in the data sets which we used in this paper. Also each record has n properties. According to the linear least square estimation method, we can write this data set in the form of an equation as shown in the formula 2 [10].

$$\sum_{j=1}^{n} X_{ij}\beta_j = y_i \qquad (i=1,2,\ldots,m) \tag{2}$$

In formula 2, β values are represent the effect coefficient of each X property. This equation can be demonstrated in a matrix format as shown in formula 3;

$$X\beta = y \tag{3}$$

The detailed notation of the above matrix is shown in formula 4.

$$X = \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}, \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \tag{4}$$

If the most converging values to the present β coefficients are represented by $\hat{\beta}$, the $\hat{\beta}$ values can be obtained by the Formula 5.

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{5}$$

By using the data within the data set which has no missing values, $\hat{\beta}$ values are obtained. In formula 3, the $\hat{\beta}$ vector is used instead of β and missing values are obtained with this way.

3.2 Expectation Maximization Method

Expectation maximization (EM) method came about for the obtainment of statistical approaches for equations that are not solvable directly [11]. It is also used for many other problems as well such as obtaining the missing values within the data. The EM algorithm is an iterative algorithm and is considered as an extension of the k-means algorithm. It carries out operations by the logic that each object is appointed to a defined set in accordance with a defined probability distribution and is consistent of two basic steps. These steps can be listed as the Expectation step and the Maximization step [12].

Let assume that there are m numbers of non-observed data in accordance with a statistical model and that Z numbers of missing values are present within this data. Let θ be a vector which shows the unknown parameters belonging to these missing values. In this case, the likelihood function is calculated as shown in formula 6 and the estimate (maximum likelihood estimate) of the θ parameters are calculated as shown in formula 7 [13].

$$L(\theta; m, Z) = P(m, Z|\theta) \tag{6}$$

$$L(\theta; m) = P(\mathrm{m}|\theta) = \textstyle\sum_Z P(m, Z|\theta) \tag{7}$$

The EM algorithm operates by using the likelihood function iteratively during the expectation and maximization steps. The formulas of the expectation and maximization steps are shown as formula 8 and formula 9, respectively [13].

$$Q(\theta|\theta^{(t)}) = E_{Z|m,\theta^{(t)}}[\log L(\theta; m, Z)] \tag{8}$$

$$\theta^{(t+1)} = arg_\theta \max Q(\theta|\theta^{(t)}) \tag{9}$$

In the expectation step, the expected value of the likelihood function is calculated by using the estimated θ parameters. The initial values of the θ parameters are defined randomly. In the maximization step, it is tried to find the θ parameters that maximize the expectation value shown in formula 8. The steps of the EM algorithm are given below [14].

1. Initialize the parameters θ to some random values.
2. Compute the best value for Z given these parameters.
3. Compute better estimate for the parameters θ with using just-computed values of Z.
4. Iterate steps 2 and 3 until convergence.

The maximum iteration number was defined as 25 during the application of the EM algorithm within the context of the study. Since the EM algorithm is an iterative method, it is fast but may not always find the ideal solution.

3.3 Hot Deck Method

The Hot Deck method is a method that finds missing data by using similar data records. All data presented in the data set is divided into groups in accordance with their similarities. A random observation is chosen from the group where the data, on which missing data appointment is to about to be made, is located. The chosen value in the observation is appointed to the missing data. Creating higher numbers of subgroups increases the validity of the estimates performed as a result of the appointment, however the not so low number of example data present in the subgroups may be a problem in this case.

Many algorithms are present for the application of the Hot Deck method. One of the most known algorithms among these is the k nearest neighbor algorithm. The most important step of this method is the determination of the distances between the data. Many methods developed for this purpose are present. The Euclidean Distance [15] method was used in this study in order to calculate the distances. Let X_i and X_j be two data contained in the data set. This data is a row vector and the number of components will be determined as the number of the properties contained in the data set. If k numbers of properties are assumed to be present in the data set, the Euclidean Distance between this data can be calculated as shown formula 10 [15].

$$d(X_i, X_j) = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{ik} - x_{jk})^2} \tag{10}$$

Calculation of the Euclidian distance is the first step of the k-NN algorithm. The process of obtaining missing values via utilization of the k-NN algorithm is given below [16];

1. Compute the Euclidean distance from the dataset.
2. Order the calculated distance by increasing order.
3. Compute k nearest neighbors, using RMSE method.
4. Calculate the average of missing k attribute that belong the missing attribute.
5. Impute the calculated average instead of the missing value.

In the study, appropriate k values (number of neighbors) were determined for the obtained different data sets and the effect of the data sets on the value k were defined as well. In the next section, the obtainment process of the data sets and the network structure are explained in detail.

4. GENERATION OF NETWORK DATA

The data sets used in the study were created in accordance with a distributed network structure. OPNET simulation tool was used to create a network data. Established network model is shown in Figure 1.

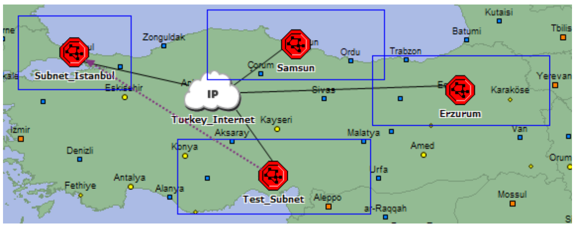


Figure 1: The network architecture to obtain datasets

Four different subnets are present in Figure 1. Among these subnets, those located in Samsun and Erzurum were created in order to model daily traffic, the subnet in Istanbul was created for modelling of the main subnet and the test subnet was created for collecting data regarding the state of the network. The IP cloud located in the center of the architecture has been used for the modelling of internet traffic present within Turkey. Three different background traffic models were added to the link between the IP cloud and the Subnet in Istanbul in order to be able to model the network traffic of the main server [17]. This traffic represents the following; the first one represents the condition of fully utilized link capacity, the second one represents package delivery at fifty percent utilization of link and the last one represents the condition where the link is almost completely empty. These three conditions were applied orderly in accordance with the Round Robin logic.

Ping packets were sent to the server located in Istanbul via a proxy server located in the test subnet. Values of the ping packet load, response time, retransmission count and the values of the throughput parameters were recorded. Data was collected under different network traffic loads under consideration of the probable different load conditions of distributed network architecture throughout the day.

Four data sets of different sizes were obtained for the created network. The size and the missing data ratios of these data sets are given in Table 1. It can be inferred from Table 1 that the loss ratio in the data increases as the size of the data increases. The reason for this is that some data cannot be obtained in frequent periods due to the simulation environment when samples are obtained more frequently

Table 1: The properties of datasets

Datasets	Number of Records	Missing Ratio
Dataset1	886	% 5
Dataset2	2360	% 18
Dataset3	7415	% 42
Dataset4	29446	% 57

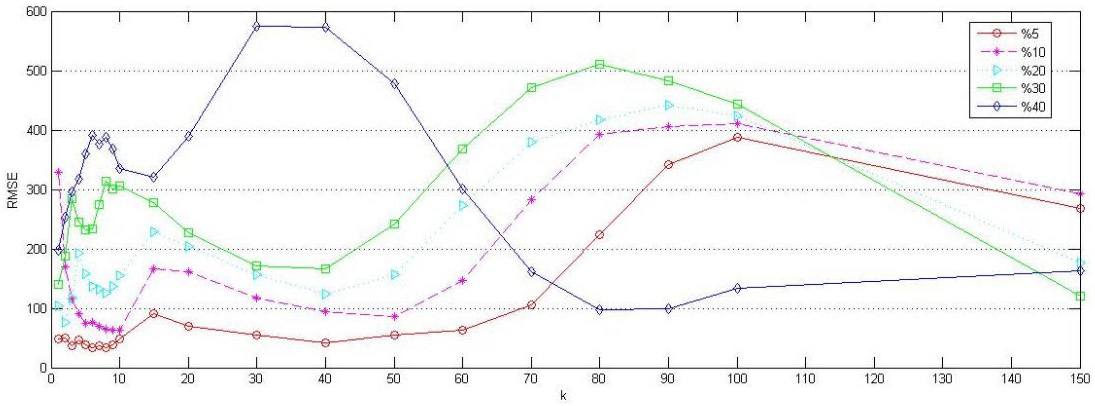
If the data sets are investigated, it can be observed that the missing values are belonging to the retransmission counts and delay parameters. The analyses were conducted on four different data sets and the effect of the size of the data set was also investigated following the success of the methods.

5. SIMULATION RESULTS

The study established to obtain missing data on four different data sets. The k-NN, LSE and EM methods were used for this purpose. First, the appropriate number of neighbors was tried to be determined via the k-NN method. After this, new data sets were obtained in such a way that they have different loss ratios than the Dataset1 and Dataset2 which were used to create these sets. The appropriate k values for the Dataset1 data set versions with %5, %10, %20, %30 and %40 missing data ratios are shown in Figure 2. The root mean square error (RMSE) function was used for the determination of the k value. The RMSE function is given in formula 11.

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M} y_{mean}^{m} - y_{imput}^{m}} \tag{11}$$

The M value in the formula represents the number of samples containing missing data, the value y_{mean}^{m} represents the average of the obtained missing values of a specific property and the value y_{imput}^{m} represents the estimated value to be placed as substitution for the missing values. The RMSE values obtained for different missing value ratios are shown in Figure 2. It can be observed that the appropriate k value is between the range 5-10 for data sets containing lost data at the rates of 5%, 10%, 20% and 30%. However, the lowest RMSE value is observed to be in 80 neighbors as soon as the loss rate of the data set reaches 40%.



,,

Figure 2: Interaction between missing ratio and k value for Dataset1

The appropriate k value for the 20%, 30%, 40% and 50% loss ratio variations of the Dataset2 are shown in Figure 3. The results are similar to the results in Figure 2. While the appropriate k value for low loss rates is between 5-10, the k value increases as the ratio of missing data increases.

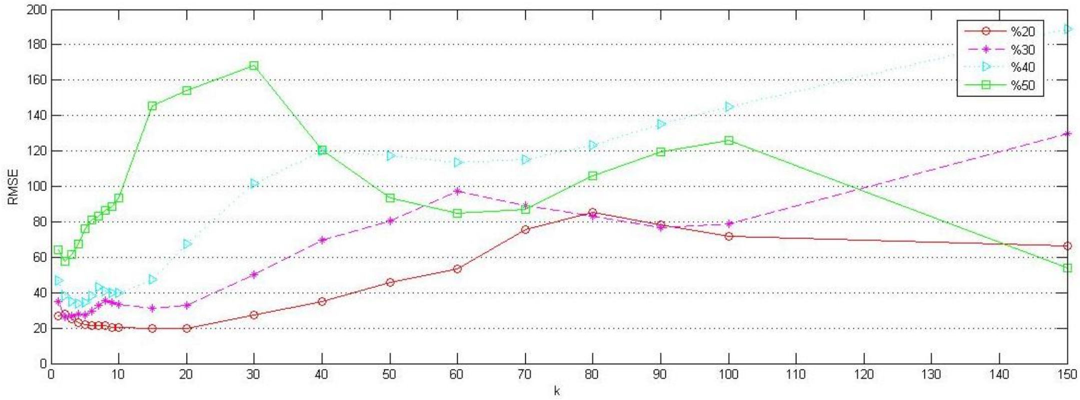


Figure 3: Interaction between missing ratio and k value for Dataset2

Afterwards, the appropriate k values for the original versions of the data sets which are shown in Table 1 were determined. The RMSE values obtained for different k values are shown in Table 2. From the values in Table 2 it can be inferred that the appropriate k values for Dataset1, Dataset2 and Dataset3 show similarity, but the appropriate k values for Dataset4 which has a loss rate of more than 50% differs from the others.

Table 2: RMSE values for different datasets

	Dataset1	Dataset2	Dataset3	Dataset4
k=1	50.07	26.56	11.84	29.10
k=2	49.96	27.70	10.64	32.82
k=3	36.09	25.14	10.16	33.36
k=4	46.83	23.12	10.10	33.76
k=5	38.50	21.79	10.29	33.90
k=6	33.96	21.63	10.37	34.17
k=7	37.54	21.22	10.29	34.06
k=8	33.56	21.30	10.34	34.15
k=9	39.07	20.56	10.57	34.28
k=10	47.57	20.27	10.64	34.27
k=15	91.05	19.90	10.56	34.49
k=20	69.80	20.03	10.33	34.69
k=30	55.18	27.55	10.80	34.72
k=40	41.90	34.71	11.28	34.63
k=50	55.05	45.55	11.85	34.71
k =60	63.49	53.31	12.40	34.72
k =70	105.37	75.60	12.98	34.50
k =80	223.33	85.25	12.49	34.16
K =90	341.90	78.09	12.55	33.98
k=100	387.84	71.64	12.95	33.90
k=150	268.74	66.46	16.60	33.40

The RMSE values for the k-NN, LSE and EM methods were obtained by using the appropriate k values calculated for the k-NN method in Table 2. The obtained results are shown in Table 3.

Table 3: Comparison of imputation techniques

Dataset	Dataset1	Dataset2	Dataset3	Dataset4
k-NN	33.56	19.90	10.10	29.10
LSE	52.38	25.99	10.35	18.09
EM	52.33	10.67	9.95	17.82

If the results in Table 3 are evaluated, it can be seen that the best result especially for the data sets with high loss rates was obtained with the EM method. The methods generally have a similar error rate. However, the EM method has a lower error rate in comparison to other methods.

6. CONCLUSIONS

The loss incurred in the data sets has a significant effect on the result of the performed analysis. Therefore, the missing values have to be obtained via various methods. The missing values occurring during the obtainment of the network data were mentioned in the study and the k-NN, LSE and EM methods used for the obtainment of missing values in different sized network traffic data were used and compared as well.

According to the obtained results, the EM method was observed to be the most stable method among these three methods. The k-NN method requires additional computation power during the determination process of the appropriate k value of the data set. At the same time, it was observed that the k value increases as the loss ratio of the data set increases. Variation was observed in the k value among different sized data sets with different loss ratios. For this reason, it has been determined that defining a standard value for the k-NN method is not very accurate and if this method should be used; the k value should be defined for especially for each data set.

7. REFERENCES

[1] Giraldo, M. M., Sanchez, J. S., Traver, V. J. 2010. A comparison of techniques for handling incomplete data with a focus on attributes relevance influence. In Proceedings of the Ninth International Conference on Machine Learning and Applications.

[2] Twala, B., Cartwright, M., Shepperd, M. 2005. Comparison of various methods for handling incomplete data in software engineering database. In Proceedings of the International Symposium on Empirical Software Engineering.

[3] Chang, G., Ge, T. 2011. Comparison of missing data imputation methods for traffic flow. In Proceedings of the International Conference on Transportation, Mechanical, and Electrical Engineering.

[4] Lıu, C. F., Chen, T. T., Lee, S. J. 2012. A comparison of approaches for dealing with missing values. In Proceedings of the International Conference on Machine Learning and Cybernetics.

[5] Y. Li, Z. Li, L. Li, "Missing traffic data: comparison of imputation methods", IET Intelligent Transport Systems, 2013.

[6] Chang, G., Ge, T. 2011. Comparison of missing data imputation methods for traffic data. In Proceedings of the International Conference on Transportation, Mechanical and Electrical Engineering.

[7] Yılmaz, H. 2014. Random Forests YöntemindeKayıpVeriProblemininİncelenmesiveSağlık AlanındaBirUygulama. Master Thesis. University of Eskişehir Osmangazi.

[8] Sezgin, E., Çelik, Y. 2013. Veri madenciliğinde kayıp veriler için kullanılan yöntemlerin karşılaştırılması. In Proceedings of the Akademik Bilişim Konferansı.

[9] Wasito, I. 2003. Least Squares Algorithms with Nearest Neighbour Techniques for Imputing Missing Data Values. Doctora Thesis. University of London.

[10] Goldberger, A. S. 1964 Econometric Theory. New York: John Wiley & Sons.

[11] C. F. J. Wu, "On the convergence properties of the EM Algorithm", The Annals of Statistics, 1983.

[12] Liu, C., Chen, T., Lee, S. 2012. A comparison of approaches for dealing with missing values. In Proceedings of the International Conference on Machine Learning and Cybernetics.

[13] A. P. Dempster, N. M. Laird, "Maximum likelihood from incomplete data via the EM Algorithm", Journal of the Royal Statistical Society, 1977.

[14] Xu, G, Zong, Y., Yang, Z. 2013 Applied Data Mining. CRC Press.

[15] Anton H., 1994. Elementary Linear Algebra. . New York: John Wiley & Sons.

[16] S. Ananthi, S. Sathyabama, "Spam Filtering Using K – nn, Journal of Computer Applications", 2009.

[17] T. Eylen, C. F. Bazlamaçcı, "One - way active delay measurement with error bounds ", IEEE Transactions on Instrumentation and Measurement, 2015.