IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 6, Ver. III (Nov.- Dec. 2017), PP 08-12
www.iosrjournals.org

Imputation of Missing Data for Network Intrusion Detection

Mehmet Ali Altuncu[1*],Fidan Kaya Gülağız[1],Tarık Bir[1],
Hikmetcan Özcan[1,] Suhap Şahin[1]
[1](*Computer Engineering Department, Kocaeli University, Turkey)*
Corresponding Author: Fidan Kaya Gülağız1

Abstract : *In data mining, occurrence of missing values in the data to be analyzed is a common issue. Ignoring these missing values, which arise from various reasons, results in failure to perform the analyses on the data in an accurate manner. Therefore, in data mining, it is necessary to identify and correct the missing values before conducting the analyses. Many methods have been developed for this purpose. In this study, synthetic missing values at various ratios were generated on the KDD Cup 99 dataset and Least Square, Naive Bayes, Hot Deck Imputation and Most Frequent Value methods were used to recover those missing values.*

Keywords: *hot deck, least square, missing value, naive bayes, network data.*

Date of Submission: 21-11-2017 Date of acceptance: 30-11-2017

I. Introduction

Missing values show up when data are missing in a survey, due to various reasons such as human errors or hardware failures. To be able to process the data accurately, these missing values in the datasets have to be replaced through the use of appropriate methods. Many studies have been conducted in various areas on obtaining the missing values. One of the most important areas that utilize these techniques is the network intrusion detection area, which works with network data. When network data are retrieved, the occurrence of missing values is common and these missing values may have a significant effect on the results to be obtained. Many methods are recommended to obtain the missing values in the literature. Whereas methods like deletion, mean imputation, and hot decking are suggested as approaches used typically for more generic datasets; techniques relying on statistical methods are utilized mainly to manage the missing values on time series data [1]. For instance, in their studies, Kaya Gülağız et al. [2] compared k-NN (k-nearest neighbors), LSE (Least Square Estimation) and EM (Expectation Maximization) methods to replace the missing values in network traffic data and observed that the EM method is the most consistent among these three approaches. Razavi-Far and Saif [3] developed a fuzzy- neighborhood density based clustering technique to eliminate the missing values. The suggested technique is to group similar patterns and to use density measurement to find out the best donors. When results are analyzed, it is concluded that the recommended technique performed better than these three techniques (k-means, fuzzy c-means (FCM) and fuzzy c-means with genetic algorithm) for obtaining the missing values. Sarıkaş et al. [4] evaluated the success of the 5 popular methods (k-nearest neighbors, Bayesian principal component analysis, local least squares, mean and median methods) to check and correct the missing values and used the NRMSE (Normalized Root Mean Square Error) parameter to measure the success rate. When NRMSE values of the methods are analyzed, it is concluded that the local least squares and Bayesian principal component analysis techniques achieved much better results compared to the others.

In this study, Least Square, Naive Bayes, Hot Deck Imputation and Most Frequent Value methods are used to obtain the missing values in the dataset, and the results are presented in a comparative analysis. KDD Cup 99 dataset is used in the study. Artificial data subsets with various ratios of missing values are obtained from the dataset and they were tested on and compared with datasets generated through models coded with the Python programming language.

II. Imputation Methods

Imputation Methods used to replace the missing data are grouped into two categories as supervised and unsupervised. Simple imputation, hot deck and cold deck imputation are classified among the unsupervised imputation methods. Predictive mean matching, regression imputation and other mechanisms based on machine learning are considered among supervised methods [5]. Even though the methods within the Simple imputation category (like mean/mode imputation) are widely preferred in many areas, they are usually incapable of resolving the missing value problem [6]. In the Hot Deck category, the missing y value is obtained through the use of samples from the survey data that have no loss in the y value. Cold Deck imputation method, which is the alternative to the Hot Deck method, has a similar logic, but needs to use a data source different from the current dataset [7]. Predictive mean matching is a regression and hot deck based method, using classification labels in

addition [5]. This way, it allows the utilization of both the supervised and unsupervised mechanisms under the same method. Methods under regression imputation obtain missing values through the use of numerical and categorical variables. Although these methods give good results with numerical data, where the variables are relevant to each other, desired outcomes may not be attained when the preferred regression model is not suitable [8].

Within the scope of the study, both supervised and unsupervised methods are used to replace the missing values in the network data. For this aim, within the unsupervised category, least square imputation and hot deck imputation among the regression imputation methods, and imputation with most frequent value among the simple imputation methods, and from the supervised category, Navie Bayes imputation method is utilized. Detailed descriptions of the methods used are given in the relevant subsections.

2.1. Least Squares Estimation (LSE)

This method aims to replace the missing value by use of the Least Squares method through certain matrix operations. The method calculates the coefficient values of each feature in the dataset using equation 1. In equation 1, *X* variable represents the two-dimensional matrix containing the properties of the two-dimensional data without the class labels, *y* variable represents the column matrix consisting of the class labels in the dataset, and beta variable is a row matrix representing the coefficient of influence for each attribute.

$$\beta' = (X^T X)^{-1} . X^T y \quad (1)$$

In order to minimize the sum of quadratic differences between existing data inputs and data reconstructed through bilinear modeling, it works in sequential order [11]. Assuming that there are *s* missing values in the dataset, the basic steps using Least Squares Data Imputation are as follows (Fig.1).

```
Set
n : number of rows in data set.
k : number of attributes in data set
s : number of missing values
for j: 1 to k
   for i : 1 to n
      Convert non-numeric features to numbers.
   end
end
Use equation 1 to obtain ß coefficients.
for i : 1 to s
   Obtain the missing value using ß.X = y.
   Write the resulting X value to the corresponding place in the dataset.
end
```

Figure 1. Pseudo code of least square estimation

2.2. Hot Deck Imputation

In this method, the Euclidean distance of each row, to the row with the missing value is calculated first, by using equation (2). Then these calculated distance values are sorted in descending order and the most frequent element among the highest *k* element is written in the place of the missing element. The process steps of the Hot Deck Imputation method are shown in Figure 2:

```
Set
k = 20;
n : number of rows containing missing data in the dataset.
m: number of rows that do not contain missing data in the dataset.
for i : 1 to n
   for i : 1 to n
      Calculate the Euclidean distance from the line to the missing line (Equation 2).
   end
   Sort the Euclidean distances from small to large.
   List the k items with the lowest distance.
   Write the most common value of the list instead of the missing value.
end
```

Figure 2. Pseudo code of hot deck imputation

In equation (2), *a* and *b* represent two different points, D(*a*, *b*) represents the Euclidean distance of these points. Also, *l* represents the dimension of these points.

$$D(a,b) = \sqrt{\sum_{i=1}^{l} (b_i - a_i)^2} \tag{2}$$

Although Hot Deck is an effective method for big data, it performs poorly with small size data. Its poor performance in cases where data are not relevant to each other is another disadvantage of the Hot Deck Method [12].

2.3. Imputation with Most Frequent Element

When compared with the other methods, the success ratio of this method is lower than the others. It replaces the most frequent element as the missing value, regardless of the other data. It is a fast technique but has high possibility of error.

2.4. Naive Bayes Imputation

Naive Bayes Imputation is used to estimate the most likely value for each missing value, with consideration of the findings provided by the surveys of each class. This operation is repeated for each attribute with missing values. If there are more than 2 attributes with missing values for a given sample, the missing value is obtained by using the attribute values with no missing values [13]. Probabilities are calculated by sorting attributes by means of relative frequencies (Equation 3). The steps of the Naive Bayes Imputation method are illustrated in Figure 3.

$$P(B_1 \mid A) = \frac{P(B_1 \cap A)}{P(A)} = \frac{P(B_1)P(A \mid B_1)}{\sum_{i=1}^{n} P(B_i)P(A \mid B_i)} \tag{3}$$

In equation 3: P(*A*) refers to the prior probability of *A*, P(*A*|*B*) refers to the conditional probability of *A* for *B*, P(*B*|*A*) refers to the conditional probability of *B* for *A*, and P(*B*) refers to the prior probability of *B* occurrence.

Set
s : number of rows in data set.
c : number of classes in data set.
k : number of attributes in data set
for *i* : 1 **to** *c*
  for *j* : 1 **to** *k*
    Calculate frequency of each feature according to class label.
  end
  Create a probability table of each class.
  Calculate conditional probability for each class using equation (3).
end
for *i* : 1 **to** *s*
  Write the value of the greatest possibility in place of the missing value.
end

Figure 3: Pseudo code of Naive Bayes imputation

III. Experimental Results

In this study, synthetic missing values are generated at various ratios on the KDD Cup 99 dataset, and four different imputation methods are tested to replace these missing values. KDD Cup 99 dataset was developed by Stolfo et al. for detection of network abnormalities. It consists of 4.900.000 lines of data, with 41 attributes each, labeled as normal or attack. There are four different types of attack in the dataset. These can be listed as, Denial of Service Attack (DoS), User to Root Attack (U2R), Remote to Local Attack (R2L) and Probing Attack [14]. Within the scope of the study, missing values are established on the KDD Cup 99 dataset in various ratios, randomly as 5%, 10%, 15% and 20%. For Naive Bayes, Hot Deck imputation and Most Frequent value methods, 50% of the dataset is used and approximately 10% is used for Least Square Imputation method.

Mean Square Error (MSE) method, which is one of the most common error detection methods, is used for the evaluation of the results obtained in the study. MSE value is calculated using equation (4).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \quad (4)$$

In equation 3; *y* variable refers to the real value of the missing data, f(*x*) value refers to the estimated value for this data, and *n* variable refers to the missing value number in the data.

Experimental tests were realized on a computer on which Intel(R) Core(TM) i7-4790K CPU @ 4.00 GHz, 8.00 GiB RAM, 24 GiB Bootdisk and Ubuntu 17.04 Operating system are uploaded. Phyton programming language was used to implement the algorithms.

Table 1 shows the error values calculated for Naive Bayes, Hot Deck Imputation, and Most Frequent Value methods, for data with missing value ratios of 5, 10, 15 and 20 percent respectively. When Hot Deck Imputation method is applied, the number of neighbors is set to 20. Previous studies [15] have shown that the number of neighbors can be set to 20 in the process of obtaining missing values for network data. Table 2 also shows the processing times obtained for the same methods.

Table 1. MSE values of implemented methods

		Hot Deck	Naive Bayes	Most Frequent Element
Missing Value Percentage	**%5**	10.787	5.894	9.644
	%10	11.051	7.571	9.870
	%15	8.529	5.633	8.299
	%20	8.259	5.446	7.964

When the results shown in Table 1 and Table 2 are analyzed, it can be concluded that Navie Bayes Imputation method has the lowest error value in various missing data ratios. The increase in missing data ratio does not cause any increase in the error rates of the methods. When the methods are evaluated in terms of transaction time, it can be seen that the methods are generally faster when the missing data ratio is low. When the methods are evaluated with regards to each other, it is observed that Navie Bayes Imputation completed processing in the shortest time.

Table 2. Process times of implemented methods

		Hot Deck	Naive Bayes	Most Frequent Element
Missing Value Percentage	**%5**	15:13:00	00:15:49	00:09:54
	%10	24:25:04	00:17:44	00:17:42
	%15	18:07:06	00:17:14	00:18:55
	%20	23:02:26	00:22:09	00:25:07

When calculations regarding LSE method were performed in Phyton language, only 10% of the data could be tested because of some problems experienced in Phyton's numPy library. The results of LSE method are shown in Table 3. Analysis of the results points out very high error values despite the use of a small ratio of the data. It is concluded that LSE method is not feasible for missing value replacement on the KDD Cup 99 dataset.

Table 3. Results of the LSE method

		MSE	Process Time
Missing Value Percentage	**%5**	618141982.638	00:00:04
	%10	16667210.429	00:00:03
	%15	456885.582	00:00:03
	%20	21941365.400	00:00:03

IV. Conclusion

In order for data to be processed in a reliable way, it is imperative to obtain the missing values in the data sets if present, through the use of the appropriate methods. Many different studies have been conducted in various areas to obtain missing values. This study focused on replacing the missing values in the data collected to detect attacks on the network.

Least square, Naive Bayes, Hot Deck imputation and Most Frequent value methods are implemented in Phyton environment to obtain the missing values that are generated synthetically with various ratios in the

network data, and their comparative performance results are presented. The analysis of the results reveals that Naive Bayes method gives more accurate and faster results in obtaining the missing values. Additionally, it is clearly demonstrated that LSE method is not feasible for replacing missing values on the KDDCup 99 dataset.

References

[1]. I. Pratama., A.E. Permanasari, I. Ardiyanto and R. Indrayani, A review of missing values handling methods on time-series data, Information Technology Systems and Innovation, 2016, 1-6.

[2]. F. Kaya Gülağız, O. Gök and A. Kavak, A Comparison of Imputation Techniques using Network Traffic Data, International Journal of Computer Applications, 142(7), 2016, 25-29.

[3]. R. Razavi-Far and M. Saif, Imputation of missing data using fuzzy neighborhood density-based clustering, International Conference on Fuzzy Systems, 2016, 1834-1841.

[4]. A. Sarıkaş, N. Odabaşıoğlu and G. Altay, Comparison of estimation methods for missing value imputation of gene expression data, Medical Technologies National Congress (TIPTEKNO), 2016, 1-4.

[5]. B. Suthar, H. Patel and A. Goswami, A survey: classification of imputation methods in data mining, International Journal of Emerging Technology and Advanced Engineering, 2(1), 2012, 309-12.

[6]. G. B. Durrant, Imputation methods for handling item-nonresponse in the social sciences: a methodological review, National Center for Research Methods Working Paper, 2005, 2.

[7]. G. E. Batista, and M.C. Monard, An analysis of four missing data treatment methods for supervised learning, Applied artificial intelligence, 17(5-6), 2003, 519-533.

[8]. I. Rozora and N. Rozora, Application of Imputation Methods for Sampling Estimation.

[9]. M. Tabassian, M. Alessandrini, R. Jasaityte, L. De Marchi, G. Masetti and J. D'hooge, Handling missing strain (rate) curves using K-nearest neighbor imputation. in Ultrasonics Symposium (IUS), 2016, pp. 1-4.

[10]. F. Soltanveis and S.H. Alizadeh, Using parametric regression and KNN algorithm with missing handling for software effort prediction, in Artificial Intelligence and Robotics (IRANOPEN), 2016, pp. 77-84.

[11]. I. Wasito and B. Mirkin, Nearest neighbour approach in the least-squares data imputation algorithms, Information Sciences, 169(1), 2005, 1-25.

[12]. T. Aljuaid and S. Sasi, Intelligent imputation technique for missing values. in Advances in Computing, Communications and Informatics 2016, pp. 2441-2445.

[13]. A. J. Garcia and E. R. Hruschka, Naive Bayes as an imputation tool for classification problems, in Hybrid Intelligent Systems, 2005, pp. 3-pp.

[14]. M. Tavallaee, E. Bagheri, W. E. Lu, A.A. Ghorbani, A detailed analysis of the KDD CUP 99 data set, in Computational Intelligence for Security and Defense Applications, 2009, pp. 1-6.

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

Mehmet Ali Altuncu,"Imputation of Missing Data for Network Intrusion Detection." IOSR Journal of Computer Engineering (IOSR-JCE) , vol. 19, no. 6, 2017, pp. 08-12.