JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Missing Data Imputation to Measure Statistic for Data Mining Applications

Shahid Ali Khan[1*], Praveen Dhyani[2]

[1, 2] Department of Computer Science, Banasthali Vidyapith, Vidyapith, Rajasthan, India

[*]*Corresponding Author: shahid.ak1972@gmail.com, Tel.: +91-9897865474*

DOI: https://doi.org/10.26438/ijcse/v7i5.12151220 | Available online at: www.ijcseonline.org

Accepted: 23/May/2019, Published: 31/May/2019

Abstract— In the applications of data mining, finding a association amongst a number of datasets is an essential concern to be focused. Correlation is generally employed in a statistical tool that supports in computing the association amongst datasets. The correlation coefficient supports in determining the strength in addition to the direction amongst two datasets and generally utilized in the real-valued datasets. In huge databases, there are various fields with mixed data types, like real, nominal and ordinal possesses values of missing information. In this paper, an effort has been made for computing the correlation coefficient between real-valued and nominal-valued dataset with missing values.

***Keywords*—**Data Mining, Real-valued data, Nominal-Valued data and Missing values.

I. INTRODUCTION

A database system is a collection of data with a specific structure or schema. A database management system (DBMS) is the software used to access a database. A data model which is independent of DBMS describes the data, attributes, and relations. There are various data models; the fundamental model is the ER (entity-relationship) data model which is often viewed as independent of the DBMS. Hierarchy, Network and Relation are three basic approaches for database systems. Relational is the most popular approach based on the mathematical concept of relation from a set A to set B as a subset of AxB. RDBMS (Relational Database Management System) is based on E.F. Codd's twelve rules, Data (2000). SQL (Structured Query Language) is the non-procedural query language used by most of RDBMS. Relations in an RDBMS are the tables where rows are referred to as tuples and columns as its attributes. Relational algebra and relational calculus are equivalent approaches to perform various operations in RDBMS. In an RDBMS, queries are well defined, and outputs are usually tables aggregate of data whereas in data mining queries are not well defined and outputs are expected to be KDD objects.

Statistical inferences leading to prediction methods are based on assumptions which may not hold for the real datasets. Data mining requires not only searching the actual data but also extracting hidden information. Statisticians deal with formatted datasets whereas the data mining deals with large datasets which may not be well formatted which are based on the concepts of sampling. In sampling, a subset of the total population is considered to obtain a general model for the entire population. Data mining goes one step further where its main objective is to find patterns and information that can be directly used by the end user. Data mining is the non-trivial extraction of implicitly unknown and potentially useful information about data [1]. Data mining is a combination of several techniques and related disciplines. Data mining is an interdisciplinary task, such as Association rule mining; Sequence mining, Predictive Modelling, Clustering and Fraud Detection. Simple statistical techniques can help to provide initial inputs for further applications of other data mining techniques. However, Data Mining technology provides a user-oriented approach to discover novel and hidden patterns or information in the data. A number of books like Data Mining concepts and Techniques by [1, 2, 3, 4, 5], examine data mining procedures and other important theories relating to data mining.

While handling the databases with diverse fields, data mining expert needs to recognize possible relationship between various domains in the dataset [6]. In data mining, correlation plays an important role to measure the degree for which the data points of one domain tend to diverge with changes in the data points of another domain, called as correlation coefficient. For example, in two hypothetical domains, A and B, the database with *n* records defines a sequence of data points, A = ($a_{1,}a_{2,}$ $a_{3,}$……..a_n) for the first domain and a sequence of values B = ($b_{1,}$ $b_{2,}$ $b_{3,}$……..b_n) for the second domain. The value of correlation coefficient, usually denoted by $r_{(x,y)}$, lying between -1 and 1, which measures the degree as well as the direction of correlation. In general, correlation is applied over two sequences of real data points [20].

The missing data in databases has been a common problem in the real world and it has become an immense challenge in the field of data mining as well machine learning [7].

Statistical analysis can be adversely affected due to missing values, particularly when there are numerous variables or when some variables are systematically missing for some units, especially in a characteristic dataset [8]. Data can be missed in a number of ways. For instance, in a survey, people usually have a tendency to leave some fields blank, or sometimes people do not have information to answer the questions. Further, some data might be accidentally lost at the time of data collection from various sources. Whatever be the causes, missing data has become a worldwide problem in the databases. Almost all standard statistical methods work on the premise that the problem, under consideration, has all the information on all the variables that need to be analyzed. In more precise terms, there can be different types of missing data, some of them are mentioned below:

A. Missing Completely At Random (MCAR)

Missing Completely At Random (MCAR) refers to that data in which the missed data does not depend upon variables, which are analyzed in the dataset [9]. Here, the data are collected and observed randomly without any regard for any variable in the dataset. However, such type of data is not often to be found and the best method to handle such data is to ignore it.

B. Missing at Random

"When given the variables X and Y, the probability of response depends on X but not on Y" [10]. This study categorizes data based on a complete and incomplete set of data. Statistical Services of University of Texas [11] suggested that it is possible to predict missing data pattern from another variable available in the database rather than focusing on a particular variable on which the data are missing.

For example:

Table 1: Missing at Random

	Variables					
Unit	1	2	3	4	5	6
1	2	5	7.5	5.5	1	4.8
2	2	-	7.5	-	-	4.8

Here both units have the same values where these observed values 2, 4 and 5 from unit 2 have the same distribution as the variables 2, 4 and 5 of unit 1.

C. Not Missing at Random

There are situations in which data is not missing at random, commonly known as Not Missing at Random (NMAR). Such a condition may occur when the missingness depends upon the real value of missing data. Modelling of such type of data is quite difficult and time-consuming task. Estimation of parameters in a data with NMAR problem is carried out by developing mathematical models of missingness.

D. Dealing with Missing Data

Two common methodologies for treating missing data in applied research are

1. to exclude the individual with missing observations from the statistical analysis or
2. to estimate the missing values and use the estimated values in the analysis. Various methods to calculate an estimate of the missing values have been suggested like, Case substitution, Hot-deck imputation, Mean substitution, Expectation-Maximization (EM), Regression imputation and Multiple Imputations.

Section I contains the introduction about data mining, data types and types of missing values. Section II covers the brief literature review of data mining applications. Section III covers the methodology used. Section IV contains the results and discussion. Section V contains the conclusion and the future scope.

II. LITERATURE REVIEW

In last few decades, several data mining techniques have been used by researcher to examine small or large data sets in different areas, like science, engineering, marketing and management to take out hidden information. For instance, in the area of marketing, it is used for recognizing customer buying behavior. The same could further be employed for customer satisfaction. Banks are used it to see the patterns associated to client credit payments trends, or loan payments etc. Data mining is also applied in biological sciences. Data mining has been used to reveal interesting information associated to it using classification as well as to find interesting and significant information from surgery databases to predict the duration of different surgeries [12].

Spiliopoulou and Pohle [13] studied the analysis and improvement of website success through data mining. For some organizations, competitiveness in web-based business requires an effective nearness on the web. Sites are utilized to build up the organization's image, to advance and offer goods and to give client bolster. The achievement of a site influences and reflects the accomplishment of the organization in the electronic market directly. This study proposed a method to enhance the "success" of sites, depending on the exploitation of navigation pattern discovery. Specifically, a theory is presented, in which achievement is demonstrated on the premise of the navigation behaviour of the site's users. We then adventure web usage mining (WUM), a navigation pattern discovery mineworker, to study how the achievement of a site is reflected in the user's conduct. With WUM the achievements of a site's parts are measured and get solid indications of how the site ought to be improved. This study provided details regarding the first methods with an online index, the accomplishment of which are considered. The mining investigation has demonstrated extremely encouraging outcomes, on the premise of which the site is currently experiencing solid enhancements.

A number of data mining techniques and their applications have been reported during the previous decade. [14] reviewed data mining techniques and their applications, and improvement they have gone through as in overview of the literature and the organization of relevant articles published from 2000 to 2011. Keyword lists and article abstracts were employed to recognize two hundreds sixteen articles concerning Data Mining Techniques (DMT) applications, from 159 scholarly journals (picked up from five online databases), this study reviews and orders DMT, as for the accompanying three areas: learning types, investigation types, and design types, together with their applications in various research and pragmatic areas. The discussion deals with any future advancement in DMT approaches and applications: DMT is resulting expanding applications in skill introduction, and the improvement of utilization for DMT is an issue-oriented domain. It is proposed that diverse sociology systems, for example, brain research, psychological science, and human behaviour may execute DMT, as another option to the existing methodologies. The capacity to ceaselessly change and gain new comprehension is the main thrust for the use of DMT, and this will permit numerous new future applications.

Data loss is normally encountered problem in social network evaluation [15]. This issue has been extensively examined in a layer network that considers one network per time. However, a natural view of multiple networks selected the problem that has not studied precisely, and findings of one layer networks were used again with no change. This research considered a detailed with the proper approach to knowing the missing data effect in complex layer networks. Based on layer dependability, ordinary network characters could either increase or decrease with regard to the complete network. The next significant aspect of experiments on fact datasets was that of multilayer network properties include layer correlation. Their significance could be employed to understand the influence of lost data over traditional network measures.

Honaker and King [20] said that utilization of modern techniques for investigating information with missing values, depends fundamentally on different attribution, have in the last half-decade get to be distinctly regular in American legislative issues and political conduct. Researchers in this subset of political science have consequently progressively kept away from the predispositions and inefficiencies brought on by ad hoc techniques like list-wise deletion and best guess ascription. However, scientists in much of near legislative issues and universal relations, and others with similar information have been not able to do the same due to the best accessible imputation methods work inadequately with the time-series cross-section information structures basic in these fields. The present study attempts to correct this circumstance with three related advancements. To begin with, fabricate a numerous imputation model that permits smooth time patterns to move crosswise over cross-sectional units, and relationships after some time and space, bringing about much more precise imputations. Second, empower investigators to incorporate learning from zone thinks about specialists through priors on individual missing cell values, as opposed to on difficult-to-interpret display parameters. Third, these undertakings could not be accomplished within existing imputation calculations, in that they can't deal with the same number of factors as required even in the easier cross-sectional information for which they were outlined, in addition, a new algorithm is developed that generously grows the scope of computationally feasible information types and sizes for which various imputation can be utilized. These improvements additionally make it conceivable to actualize the techniques presented herein unreservedly accessible open source software that is impressively more consistent than existing calculations.

In data mining, correlation plays an important role. The data mining experts could use information obtained after measuring the relationship for further analysis. For example, information received from the association could be employed in regression analysis. A common multivariate technique like principal component analysis (PCA) established on correlation analysis [17, 20] and further correlation was being employed to calculate the distance among data points in clustering [18, 20]. These two systems were employed in handling data with real values only. For nominal-valued data, [19] proposed d_{cv} metric that was similar to Mahalanobis metric based on Cramer's V-statistics.

An interesting study by Cabana states that "about 20% of effort is spent on the problem and data understanding, about 60% on data preparation and about 20% on data mining and analysis of knowledge". People spend more time on data preparation? Because, there were many serious data quality issues in real datasets, like incompleteness, redundancy, inconsistency and noise. These problems reduce the performance of the data mining algorithm. Missing data is a common problem in almost every real dataset. But, if the rate of missing is less than 1%, missing data may not make significant difference for the knowledge discovery in database process, 1-5% is manageable, "5-15% requires sophisticated methods to handle, while more than 15% might severely impact any kind of interpretation". Though, sometimes, a missing value in the database specifies that the value is zero or in certain other cases it can't probably exist. For example, an entry in a field of salary for a child. Though, in several cases, an empty field might represent an unidentified quantity and, in such cases,, applications, depends on correlation, could be utilized to complete that entry.

Data Mining field values are of different data types that embrace nominal, ordinal or real. Methods for computing correlation between the sequences of mixed data were studied regardless of their types. Specifically, an innovative method for finding the correlation among nominal and real valued data were suggested by [6]. The technique depends on the assigned nominal values and therefore, it is known as A-correlation (A for assignment). The planned work is the most favorable for assignments which could be employed for better computation. Cramer's V-statistic could be employed even for ordinal case. In case of the ordinal data, correlations depending on rank were suitable and these might be utilized in the ordinal and real cases.

In the applications of data mining, measurement of relationship amongst a number of datasets is an essential concern to be focused. Correlation is generally employed in a statistical tool that supports in computing the relationship amongst datasets. The correlation coefficient supports in determining the strength in addition to the direction amongst two datasets and generally utilized in the real-valued datasets background. In huge databases, there are various fields with mixed data types, like real, nominal and ordinal possesses values of missing information. In this paper, an effort has been made for computing the correlation coefficient between real-valued and nominal-valued datasets with missing values [20].

III. Methodology

When one of the sequences is real-valued, and the other one is nominal-valued, i.e., there is no ordering of elements, the correlation between them could be measured as follows. Now consider a sequence of n real-valued data, Y**,** and a 'n' nominal values sequence, Z, that could be carefully chosen from a particular domain K = { $v_1, v_2, \ldots\ldots\ldots v_q$}
Now through every v_k, $1 \leq k \leq q$, here would be a related set of indices

$$J_k = \{ j \mid 1 \leq j \leq n\_\text{and}\_z_j = v_k \}$$

Let us assume that y^k that is constructed from **Y** denotes the sequence of real values by selecting the *y*-values with indices in J_k. Hence y^k includes y-values. The *y*-values contain the corresponding ***z***-values that would be equal to V_k.
Let n_k = $|J_k|$ that be the length of Y^k and $\overline{y_k} = \sum_{j \in J_k} y_j / n_k$ which represents the mean of the values in $\mathbf{Y}^k$.

Now define

$$T_{yy}^k = \sum_{j \in J_k} (\overline{y_k} - y_j)^2$$

$$Corr_V(\text{y,z}) = \sum_{j \in J_k} (\overline{y}_k^2 - 2\overline{y}_k y_j + y_j^2)$$

$$= \sum_{j \in J_k} y_j^2 - n_k \overline{y}_k^2 \qquad (1)$$

And hence the standard deviation of y^k is.

$$\sigma_{y^k} = \sqrt{\frac{T_{yy}^k}{n_k}} \qquad (2)$$

Relating the differences of weighted average of Y^k, $1 \leq k \leq q$ accompanied by non-zero variance of **Y** would be utilized to determine real-valued sequence correlation 'Y' in addition to the sequence of nominal-valued Z. This is a simple statistic to measure the correlation between real and nominal valued sequences [20].

This stimulated the definition:

$$corr_v(y,z) = \frac{T_{yy} - \sum_{k=1}^{q} T_{yy}^k}{T_{yy}}$$

$$= \frac{(\sum_{j=1}^{n} y_j^2 - n\overline{y}^2) - \sum_{k=1}^{q} (\sum_{j \in J_k} y_j^2 - n_k \overline{y}_k^2)}{(\sum_{j=1}^{n} y_j^2 - n\overline{y}^2)}$$

$$= \frac{\sum_{k=1}^{q} n_k \overline{y}_k^2 - n\overline{y}^2}{\sum_{j=1}^{n} y_j^2 - n\overline{y}^2} \qquad (3)$$

It has been noted that $corr_v\ (y\ ,z) = 1$, iff each $T_{yy}^k = 0$, i.e., every single y^k is the sequence of same values, otherwise,
$0 < corr_v\ (y,z) < 1$. The $corr_v\ (y,\ z)$ can never be attained negative, while dealing with nominal data as nominal data is characteristically random or unordered in nature.

If the size of n as well as n_k $(1 \leq k \leq q)$ is large and each of y^k be outlined arbitrarily from the values which are existing in **y** formerly σ_{y^k} would be a good approximation of σ_y. So, as estimated, By using simple algebra, it can easily be confirmed that for

any $\lambda \neq 0$ and μ,

$$corr_{V}(\lambda y + \mu, z) = corr_{V}(y, z) \qquad (4)$$

Thus, $Corr_V$(y, z) have the distinct, unique features that everyone could able to predict correlation-measure.

Another parallel approach is concerning to construct initial a map of *V* onto *R* at that point relate this mapping to the sequence of z to transform all members into the sequence of real valued . In view of the fact that elements associated with V belong to nominal category, they might be transformed over with any injective mapping; an advantage behind mapping to a real-valued sequence is that $U \subset R$ and when we ought to adapted z to a real-valued sequence, the correlation between y and the transformed sequence is surely recognized. There are vastly numerous such mappings $\boldsymbol{V} \longrightarrow \boldsymbol{R}$ and the best appropriate need to be deliberately picked.
For any particular α : A mapping $\boldsymbol{V} \longrightarrow \boldsymbol{R}$**,** a sequence α(z) is defined as

$$\alpha(z) = (\alpha(z_1), \alpha(z_2), \alpha(z_3), \ldots\ldots \alpha(z_n))$$

and subsequently it is feasible to measure the correlation between the sequences y and $\alpha(z)$. Assume $\boldsymbol{V} \rightarrow \boldsymbol{R}$ stand for the set of all possible mappings from V to R, at that point $Corr_V$(y, z), the assignment correlation, between y and z may be defined as,

$$corr_{A}(y,z) = corr_{A}(z,y) = \max\{corr(y,\alpha(z)) \big| \alpha \in V \longrightarrow R\}$$

At this point, α required to be injective, here, this requirement is relaxed; however, the maximizing function is generally expected to be injective, and even though it is not injective, the slight modifications is made in encrypting to achieve the property of injective and could formulate only little difference to the ensuing correlation that could be carefully uncared. Subsequently, a new real-valued sequence is constructed, $z' = \mu(z)$ where $\mu(\upsilon_k) = \overline{y_k}$ for $1 \leq k \leq q$. This intends to show the setting as $\alpha = \mu$ which maximizes *corr(y, α(z))* initially, *corr(y, z')* ensuing property is ascertained.

IV. Results and Discussion

The illustration of computing correlation between nominal and real-valued variables with and without missing values by using Benchmarked dataset from Machine Learning Database Repository at the University of California, Irvine. The data file *servodata.xls* contains three variables and 167 instances.

Table 2: Comparison of *corr*

Table 2: Comparison of $corr_v$ *(y, z)* results

	$Corr_v$ (y, z)					
	Moter-Pgain	Moter-Vgain	Moter -Class	Screw -Pgain	Screw -Vgain	Screw -Class
Complete Datasets	0.021	0.006	0.039	0.014	0.005	0.022
Mean substitution	0.023	0.005	0.037	0.021	0.006	0.023
Discarding missing values	0.026	0.007	0.042	0.026	0.008	0.029
Hot-deck	0.032	0.009	0.033	0.027	0.019	0.025

Table 3: Comparison of $corr_A$ *(y, z)* results

	$Corr_A$ (y, z)					
	Moter-Pgain	Moter-Vgain	Moter -Class	Screw -Pgain	Screw -Vgain	Screw -Class
Complete Datasets	0.145	0.077	0.197	0.118	0.071	0.148
Mean substitution	0.152	0.071	0.192	0.145	0.077	0.152
Discarding missing values	0.161	0.084	0.205	0.161	0.089	0.170
Hot-deck	0.179	0.095	0.182	0.164	0.138	0.158

The correlation coefficients $corr_V$ *(y, z)* and $corr_A$ *(y, z)* between real-valued data and nominal-valued data were computed by two different approaches and then compared the results of as shown in Table 2 and Table 3 respectively, we found that the values of the correlation coefficient, with and without missing data were closer by mean substitution. Hence, missing data imputation by exploiting mean substitution was an appropriate technique.

V. Conclusion and Future Scope

There are several serious data quality issues in real datasets like incompleteness, redundancy, inconsistency and noisy. Such complications decrease the performance of the data mining algorithm. The proposed research provided suitable techniques for replacing missing values more effectively and then computed the correlation between various types of mixed type datasets. This study will help data miners, academicians and students who face difficulty in measuring association between two datasets with and without missing data.

REFERENCES:

[1] Fayyad U M, Piatetsky-Shapiro G, and Smyth P, ”Advance in Knowledge Discovery and Data Mining” ,1-34, Menlo Park, CA:AAAI Press/MIT Press, 1996a.

[2] Berry M J A and Linoff G, “Data Mining Techniques for Marketing, Sales and Customer Support” , NY: John Wiley and Sons, 1997.

[3] Hand D, Mannila H, and Smyth P, “Principles of Data Mining”, Prentice-Hall of India Private Limited, India, 2001.

[4] Han J and Kamber M, “Data Mining: Concepts and Techniques” , San Francisco, Morgan Kauffmann Publishers, 2001.

[5] Dunham M H, “Data Mining: Introductory and Advanced Topics”, 1st Edition Pearson Education (Singapore) Pte. Ltd., 2003.

[6] Rayward-Smith, V. J., “Statistics to measure correlation for data mining applications”, *Computational Statistics & Data Analysis*, *51*(8), 3968–3982. doi:10.1016/j.csda.2006.05.025, 2007.

[7] Hong, T.P., Wu, C.W., “Mining Rules from an incomplete dataset with a high missing rate”, *Expert Systems with Applications*, *38*(4), 3931–3936. doi:10.1016/j.eswa. 2010.09.054. 2011.

[8] Ferrari P.A., Annoni P., Barbiero A., Manzi G.,” An imputation method for categorical variables with application to nonlinear principal component analysis”, *Computational Statistics & Data Analysis*, *55*(7), 2410–2420. doi:10.1016/j.csda.2011.02.007.

[9] Judi Scheffer, “Dealing with Missing Data”, *Res. Lett.Inf. Math.Sci (2002). Quad A, Massey University, P.O. Box 102904 N.S.M.C, Auckland, 1310.*

[10] Rubin D.B., “Inference and missing data”, Biometrika 63(3):581–592,1976.

[11] Statistical Services of University of Texas, “ Handling missing or incomplete data” ,2000.

[12] Combes, C., Meskens, N., Rivat, C., & Vandamme, J. P.,” Using a KDD process to forecast the duration of surgery”, *International Journal of Production Economics*, *112*(1), 279-293, 2008.

[13] Spiliopoulou, M., & Pohle, C.,” Data mining for measuring and improving the success of web sites”, *Data Mining and Knowledge Discovery*, *5*(1-2), 85-114, 2001.

[14] Liao, S. H., Chu, P. H., & Hsiao, P. Y. ,” Data mining techniques and applications–A decade review from 2000 to 2011”, *Expert Systems with Applications*, *39*(12), 11303-11311,2012.

[15] Sharma, R., Magnani, M., &Montesi, D.,” Investigating the types and effects of missing data in multilayer networks”, In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 392-399). ACM, 2015.

[16] Kossinets, G.,”Effects of missing data in social networks”, *Social networks*, *28*(3), 247-268,2006.

[17] Jolliffe, I.T. ,”Principal Component Analysis”, Springer, Berlin,1986.

[18] Jain, A.K., Murty, M.N., Flynn, P.J., “Data clustering: a review”, ACM Comput. Surveys 31 (3), 264–323,1999.

[19] Al-Harbi, S.H., McKeown, G.P., Rayward-Smith, V.J., “A innovative metric for categorical data”, In: Bozdogan, H. (Ed.), Statistical Data Mining and Applications,2003.

[20] Shahid Ali Khan, Praveen Dhyani, ”*Data Mining: Using C++ to Measure Correlation between Real-Valued and Nominal Valued datasets,*" International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN: 2347-5552, Volume-3, Issue-2, March 2015.

Authors Profile

Mr. Shahid Ali Khan pursed B.Sc.(Hons), M.Sc.(Statistics) and MCA from Aligarh Muslim University, Aligarh India in 1995, 1997 and 2001 respevtively. He got University Medal in B.Sc ang PG merit Scholarship in M,Sc. He worked as Assistant Professor at International Center Muscat, B.I.T.Mesra in the Deaprtment of Computer Science and Engineering during 2004 to 2017 out of wich 6 years served as In-Charge of examinations and currently pursuing Ph.D from Banasthali Vidyapith, Rajasthan, India. He has 15 years of teaching experience and 2 years of Research Experience.

Prof (Dr.) Praveen Dhyani, Department of Computer Science, Banasthali Vidyapith, Rajasthan, India. He has held the positions like Executive Director at Banasthali Vidyapith- Jaipur Campus. He has also worked with BIT Mesra, Ranchi and held the positions like, Rector, Director at BITIC Muscat and UAE centers. He started his career as teaching faculty in Computer Science Department of BITS, Pilani-Rajasthan from 1973 to 1987.