Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 8, August 2015, pg.481 – 485

RESEARCH ARTICLE

Intelligent Based Imputation Methods for Text Mining Applications to Phishing Attacks

Soujanya Maddina

M. Tech Student (IT), VNR VJIET, Hyderabad India

Abstract— Real world datasets commonly consist of missing values. Missing data creates serious issues in various applications like statistics, data mining and machine learning and others. Phishing attack causes financial loss. In this paper we impute the fiscal dataset based on the proposed data imputation approach. Secondly, we observe the text mining on the textual information (unstructured information) of phishing assaults. Thus, the textual data is modified to structural knowledge. In the end, we can predict the risk factor of phishing attacks based on combined data from the economic announcement of the organizations and textual data making use of classifiers.

Index Terms— Data Imputation, K Means, Phishing alerts, Text Mining

I. INTRODUCTION

In the message, or some estimate is worth to lack the ability to point the lack of capacity factor replacement area. When all the estimates missing values, this data set will be present for the entire knowledge utilization analysis. In a real understanding of the project is missing data in the presence of many disciplines in an unavoidable problem. To check the available data, the integrity and quality of information plays a fundamental role on the grounds that the inferences made by the entire information is not from these incomplete data to make the right additional [1]. For example, researchers found unusual settings all investigation. Of respondents did not intend to offer the overall advantage considering the fact that neglect of survey questions, or ambiguous said. Lack of material variable can be used to check the data of the primary things. Therefore, the main problem in this imputation data execution priority.

Imputation can also be very invaluable knowledge manipulation headquarters purposes, such as website visitors monitoring, industrial approach, telecommunications and computer networks, automatic speech recognition, financial and commercial use, as well as medical diagnosis, amongst others. Data because the data input error, system failures and noise channel missed. According to the literature [2], the lack of information is divided into three categories :(a) random completely missing: no additional information can help researchers to fill in missing values. Data decline, due to structural reasons not to be considered as the missing completely at random (II) in the absence of random: random deletion of information need to be able to fill the lack of models to provide greater support qualified variables. For example, if the lack of age-based field of qualified people to do the degree of imputation.(C) non-random missing: Learn data to estimate other variables. For example, the use of ZIP codes estimates. Lack of information to create multiple image data mining, arithmetic, many problems in the field of statistical research, as well as many other areas [1]. To blame incomplete or lack of data, a couple of statistical evaluation provisions established procedures [3]. These include alternative methods mean, hot deck imputation, regression method, EM, and a couple of estimation methods.

Phishing is a way where in person attempt to acquire information from the sufferer impersonating as a dependable third social gathering [4]. Phishing scams are growing daily. It is the important hazard in online neighborhood. And the way in which it's carried out now a day is via mails considering that mails are essentially the most customary way of verbal exchange all over the world. The way web has grown at recent instances where most of the transactions occur on-line and this certainly is the cause for growing cybercrimes. It's also degrading the belief of persons who make their lifestyles easier with the aid of using internet. An attacker can be using few procedures to get extra information. Phishers are becoming smarter. They are attempting to mix both technical and social vulnerabilities. Internet customers are victims of the attack. To get know-how, he can use n/w websites.

- One could get social information by way of mining web content and hyperlinks.
- They no longer best rationale economic loss but in addition shatter trust of folks who use e-commerce.
- Data mining methods would be used to verify the phishing assaults based on prior incidents.
- Fiscal loss is a greater concern.
- A Warning message concerning phishing attacks will likely be of higher interest.
- Patterns involving prior incidents will form a foundation for the prediction of future incidents before they happen.

Information mining systems would be used to assess the phishing assaults founded on the past incidents. Fiscal loss shall be a better crisis. For that reason, warning message regarding phishing assaults might be a greater interest. Patterns regarding prior incidents might be a ground work for prediction of future incidents earlier than they occur.

Intelligent based Imputation Methods to phishing alerts employs two stage soft computing technique to entry the severity of phishing assaults. Imputation is used to fill lacking values. Hybrid approach is used to predict the severity of Phishing attacks. For that reason we are utilizing a process in which we use k means with different distance measure to fill in lacking values along with MLP. After imputing lacking values, we mine the economic working out involving the organizations at the side of the structured potential of the textual data making use of classifiers. Total accuracies are found out.

II. RELATED WORK

Remove and estimates: the missing values can be handled in two ways. Ignore the lack of knowledge or lack of technical systems, will be removed, which would lead to inaccurate results. It is used as a default solution, but it was not accepted, because it can eliminate valuable information. This process has two forms :(a) a list of deleted simply ignore instances contain missing values. The drawback of this mechanism is that the utility can lead to a lack of large-scale observation range, which will affect and increase the long-established understanding if additional set itself too small at high error. (Ⅱ) pair deletion process to consider the role of each individual. All recorded values to view and understand the unknown omitted. This approach is entirely correct, when the total sample of small size or lack of knowledge of the case is significant. Estimation method using its already existing information. The simplest system is the mean estimate, where the value of a variable changes missing from the average value of the entire remaining information of the variable. The problem with this approach is that it ignores the considerable correlation between the components [5]. When the variables are correlated, information interpolation may be complete by regression imputation. In regression imputation, regression equation by using the watch contains an incomplete value, because the properties intention of variables, as long as the calculation. This method preserves the lack of professional knowledge and other variables variance and covariance. Hot and cold deck imputation case replace the complete lack of values and the closest, where "recent" is in add-ons, this may be other incentives for each case with missing values for each vector [5]. It requires the in Reply respondents were assigned a value basis. If some value is missing then it requested information from others, and filled value obstacle is the lack of the advantages of this process is estimated to be located in a single entire vector, thus ignoring the data set value on a global scale. important principle cold deck imputation is missing value in the data set exclusive value changes. The main problem is the cold deck imputation of missing values replaced dataset exclusive value [2].

After a couple of estimation methods, each value by a set of affordable and legitimate values for each occasion worthy of M and analyzed by the exchange, with the inference is becoming all the data gathered, we get M for the entire information unit. According to the literature [2], a couple liability than in the case, sensible and pleasant, with an average replacement estimates. Regression is less a couple of interpolation. EM is an iterative method to continue, unless there are possible parameters estimates restraint.

Another method is to use the real k- nearest neighbors (K-NN). Missing values of the k-NN method at its nearest neighbors by replacing. Nearest neighbor is from the environment, thereby reducing the gap between performance [6] [7] .Neural networks have been interpolated for information and options. Neural technology community, MLP is as proficient use of the entire case, a decision on a variable target, whenever a regression method. Text mining is one of the user by using a set of analysis tools to approach a group of files with the interaction of intensive systems. Text Mining aims to identify sources of information interesting pattern extraction [8] and Exploration.

In characteristic- text mining algorithms function creates a record indicates. Words, phrases and terminology are used to symbolize the documentation. Text mining available because of its potential mining status available on the Web unstructured / digital content material as a research facility. It tries to find important information from the available sources. It includes traditional first change unstructured content into structured content material, the use of normal data mining program before. Free text phishing alert is changed to a matrix structure in terms of the document file. Key phrase text mining methods to extract, cluster analysis and conceptual hyperlinks, observe special form of disclosure of information about security incidents in the financial statements. Text and data mining constitutes a hybrid approach for some researchers greater interest.

III. METHODLOGY

A. Data Imputation

For the data imputation, we used K means algorithm with a new similarity measure. The procedure for data imputation is as follows:

Algorithmic steps to perform k-means imputation: Let D={d_1,d_2,d_3,d_4..d_n} be the set of data points, C={c_1,c_2,c_3,c_4…c_n} be the set of cluster centers and D_m=D_{m1},D_{m2}…D_{mn} be the missing values within the data D and let k be the number of clusters.

1. Initialize cluster centers through random selection.
2. Now, calculate the distance between cluster centers and the complete data points. Distance is calculated using chebyshev distance measure.

$$\max\sum_{i=1}^{n} \left| d_i - c_i \right|$$

3. Assign the data points to the closest cluster.
4. Calculate mean for the clusters formed and update with the random cluster points which are taken in step 1 using the following formula.

$$C_i = \frac{1}{P}(\sum xi \ )$$

Where p is the number of points in that particular cluster.

5. Calculate the distance between the new cluster centers with every other point in the data and calculate mean value.
6. If the mean is equal to the mean value obtained in the previous step then stop the process else continue iterating.
7. For missing data points calculate the distance between mean values and the complete data points, of incomplete data points and fill in the values.

B. Text Mining

For the purpose of text mining data is taken from millersmiles database. The entire set of text files is converted to lowercase followed by removing delimiters, removing stop words and stemming. The entire set of documents is taken to form a document term matrix which is given as an input to the classifier.

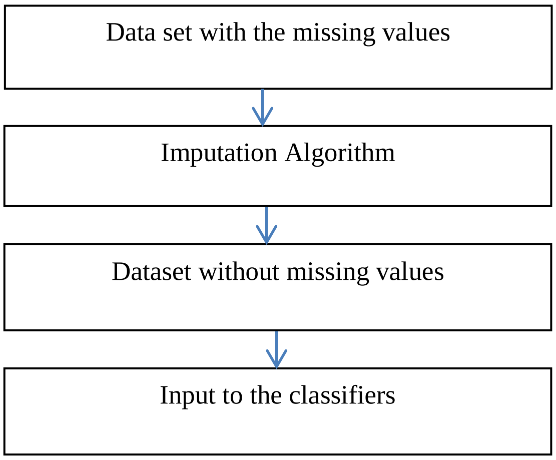
Data set with the missing values
Imputation Algorithm
Dataset without missing values
Input to the classifiers

Fig. 1 System Overview

IV. SYSTEM IMPLEMENTATION

Top 25 variables are chosen for the Classification purpose, as proposed by Chen et al. (2010). Missing values are filled using the approach as stated above. Textual data is converted to structural form and both textual data and fiscal data is given as an input to classifier.

For textual analysis variables considered are:

Account, Assets, Bank, Computer, Confirmation, Consumers, E-bay, E-mail, Information, Security, Update, Warning and Work

V.RESULTS

Before working on the actual data we tried working on different types of data sets. Results are obtained by using an open source tool. The following are the data sets which are considered for testing the algorithm taken from KEEL.

Table 1

Datasets	Attributes	Examples	Class	MValues
Bands(BAN)	19(13/6/0)	539	2	32.28
Ecoli(ECO)	7(7/0/0)	336	8	48.21
Iris(IRS)	4(4/0/0)	150	3	32.67
Pima(PIM)	8(8/0/0)	768	2	50.65
Wine(WIN)	13(13/0/0)	178	3	70.22

Table 2: Testing our algorithm with other methods using classifier (C4.5)

Datasets	Our Algorithm	KNNI	SVMI	WKNNI	KMI
BAN	70.11	70.32	69.18	69.57	70.11
ECO	82.14	82.15	82.50	82.14	81.88
IRS	96.00	96.00	96.00	96.00	96.00
PIM	75.25	71.09	73.32	73.69	72.78
WIN	87.54	87.64	86.56	88.75	86.54

VI. CONCLUSION

In this paper, we offered a novel method for imputing the missing values with a different similarity measure for k means. Classifier accuracies are found out using open source tool Weka. We applied text mining on the data of phishing alerts and converted that document to structured form. Finally, we predicted the risk factor from both the financial data and structured data.

REFERENCES

[1] Abdella,M, Marwala, D. (2005).The use of genetic algorithms and neural networks to approximate missing data in database. Proc. ICCC(pp.207-212)

[2] Little, R.J.A & Rubin, D.B (2002).Statistical analysis on missing instance(2nd edition) Willey-Interscience

[3] Garcıa-Laencina, Sancho-Gomez, & Figueiras-Vidal, 2010. Pattern Classification with missing data.A review. Neural Computing and applications,19,263-282.

[4] https://en.wikipedia.org/wiki/Phishing.

[5] Schafer, 1997.Analysis of multivariate data.

[6] Batista, G., & Monard, M. C. (2002). A study of K-nearest neighbor as an imputation method. Hybrid intelligent systems, ser front artificial intelligence applications (pp. 251–260). IOS press.

[7] Batista, G., & Monard, M. C. (2003). Experimental comparison of K-nearest neighbour and mean or mode imputation methods with the internal strategies used by C4.5 and CN2 to treat missing data.

[8] Srinivasan P (2003).Text mining. Generating hypothesis from medicine. Journal of the American Society for Information Science and Technology.

[9] Chen, X., Bose, I., Leung, A. C. M., & Guo, C. (2010). Assessing the severity of phishing attacks: A hybrid data mining approach. Decision Support Systems, 50, 662–672.

[10] Ankaiah, N., & Ravi, V. (2011). A novel soft computing hybrid for data imputation. In Proceedings of the 7th international conference on data mining (DMIN).