Expert Systems with Applications 39 (2012) 10583–10589

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts

Kancherla Jonah Nishanth [a], Vadlamani Ravi [a], Narravula Ankaiah [a], Indranil Bose [b,*]

[a] Institute for Development and Research in Banking Technology (IDRBT), Castle Hills Road #1, Masab Tank, Hyderabad-500 057, AP, India
[b] Indian Institute of Management Calcutta, Diamond Harbour Road Joka, Kolkata 700 104, West Bengal, India

ARTICLE INFO

Keywords:
Data imputation
K-means clustering
Multilayer perceptron
Phishing alerts
Probabilistic neural networks
Text mining

ABSTRACT

In this paper, we employ a novel two-stage soft computing approach for data imputation to assess the severity of phishing attacks. The imputation method involves K-means algorithm and multilayer perceptron (MLP) working in tandem. The hybrid is applied to replace the missing values of financial data which is used for predicting the severity of phishing attacks in financial firms. After imputing the missing values, we mine the financial data related to the firms along with the structured form of the textual data using multilayer perceptron (MLP), probabilistic neural network (PNN) and decision trees (DT) separately. Of particular significance is the overall classification accuracy of 81.80%, 82.58%, and 82.19% obtained using MLP, PNN, and DT respectively. It is observed that the present results outperform those of prior research. The overall classification accuracies for the three risk levels of phishing attacks using the classifiers MLP, PNN, and DT are also superior.

1. Introduction

In statistics, imputation is the substitution of some value for a missing data point or a missing component of a data point. Once all missing values have been imputed, the dataset can then be analyzed using standard techniques for complete data. Missing data in real life data sets is an unavoidable problem in many disciplines. For analyzing the available data, completeness and quality of the data plays a major role because the inferences made from complete data are more accurate than those made from incomplete data (Abdella & Marwala, 2005). For example researchers rarely find the survey data set with complete entries (Hai & Shouhong, 2010). The respondents may not give complete information because of negligence, privacy reasons, or ambiguity of the survey questions. The missing parts of variables may be important things for analyzing the data. So in this situation data imputation plays a major role. Data imputation is also very useful in the control based applications like traffic monitoring, industrial processes, telecommunications and computer networks, automatic speech recognition, financial and business applications, and medical diagnosis, among others.

Data in the databases may be missed because of data entry errors, system failures at the time of data retrieval or several other reasons like sensor failures, noisy channels, and cultural issues in updating the databases etc. According to Little and Rubin (2002), missing data is categorized into three categories: (i) missing completely at random (MCAR), (ii) missing at random (MAR), (iii) not missing at random (NMAR). MCAR occurs if the probability of missing value on some variable X is independent of the variable itself and on the values of any other variables in the dataset. For example, if the age of the husband is missing in a customer's database then it does not depend on the any other variable of database which is meant for wife. MAR occurs if the probability of missing value of some variable X is independent of the variable but the pattern of missing data can be traceable or predictable from other variables in the database. For example, if income of a person is missing, then one can predict the missing value by using the values in profession and age. NMAR occurs when the probability of missing value of some variable X depends on the variable X itself. For instance, if citizens do not participate in a survey, then NMAR occurs. MCAR and MAR data are recoverable, whereas NMAR data are irrecoverable.

Missing data creates various problems in many research areas like data mining, mathematics, statistics, and various other fields (Abdella & Marwala, 2005). To impute with incomplete or missing data, several techniques based on statistical analysis are reported (Garcıa-Laencina, Sancho-Gomez, & Figueiras-Vidal, 2010). These methods include mean substitution methods, hot deck imputation, regression methods, expectation maximization, and multiple imputation methods. Other machine learning based methods include self-organizing maps (Merlin, Sorjamaa, Maillet, & Lendasse,

* Corresponding author. Tel.: +91 33 2467 8300x157.
E-mail addresses: jonah.nishanth@gmail.com (K.J. Nishanth), rav_padma@yahoo.com (V. Ravi), ankireddy.cse@gmail.com (N. Ankaiah), indranil_bose@yahoo.com (I. Bose).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.02.138

2010), K-nearest neighbor (Batista & Monard, 2002), multi layer perceptron (Gupta & Lam, 1996), fuzzy-neural network (Gabrys, 2002), and auto-associative neural network imputation with genetic algorithms (Abdella & Marwala, 2005), among others.

1.1. Phishing attacks

Phishing is a way of attempting to acquire sensitive information such as usernames, passwords and credit card details by masquerading as a trustworthy entity in an electronic communication (http://en.wikipedia.org/wiki/Phishing). Phishing is a major security threat to the online community. Phishing scams have been escalating in number and sophistication by the day. A phishing attack today targets audience by using mass-mailings to millions of email addresses around the world, as well as by communicating with highly targeted groups of customers that have been enumerated through security faults in small clicks-and-mortar retail websites (Jagatic, Johnson, & Menczer, 2006). Phishing attacks in the US soared in 2007 as $3.2 billion was lost to these attacks, according to a survey by Gartner, Inc. The survey found that 3.6 million adults lost money in phishing attacks in the twelve months ending in August 2007, as compared to the figure of 2.3 million the year before. According to the Anti-phishing Working Group 2009 reports, the focus of phishing attacks is more on the payment services industry (37.9%) and the financial services industry (33.1%) (APWG, 2009). Not only do phishing attacks cause financial loss, but they also shatter the confidence of customers in conducting e-commerce.

Data mining techniques can improve the assessment and prediction of phishing attacks. Data mining techniques discover the patterns associated with past incidents and then predict the future incidents before they happen. The financial loss resulting from a phishing attack is always of great concern to security administrators as well as consumers of an organization. Therefore, a warning mechanism that can identify the phishing incidents that are either very risky or likely to cause a large financial loss is of great interest to shareholders and senior managers of the targeted companies. Jackobsson and Myers (2007) classified the costs of phishing into three categories namely *direct cost, indirect cost and opportunity cost*. Singh (2007) studied a number of international phishing incidents and found out that the direct financial loss per incident ranged from US $900 to $6.5 million. For filtering out phishing emails containing fraudulent message data mining was used (Airoldi & Malin, 2004). Researchers have also investigated if the factors that drive successful marketing campaigns also drive successful phishing attacks (Workman, 2008).

Chen, Bose, Leung, and Guo (2010) employed hybrid data and text mining to access the severity of phishing attacks in firms using classifiers viz., DT, SVM, and MLP. Their paper forms the background for the research reported in this paper. To determine the severity of phishing attacks two types of input data, namely, phishing alerts from the database Millersmiles and financial data available from the financial statement of the firms were utilized by Chen et al. The phishing alerts data used was the largest available phishing alerts dataset at the time of research, and was collected from mid-2005 to mid-2008. The technical sophistication of the phishing attack was measured in terms of the risk level of the attack that was determined by the information security specialists of Millersmiles. These risk levels were: low, low-medium, medium, medium-high, and high. The majority of alerts belonged to the risk level medium. For the sake of simplicity, the risk levels low and low-medium were grouped to from a new group 'low' and medium-high and high were combined to form a new group 'high'. In the raw financial dataset there were 168 attributes. Chen et al. (2010) considered the 75 attributes related to the financial performance of the firm. They used the Pearson chi-square statistic to determine the strength of relationship between 75 financial variables and the target variable, i.e., risk level. The top 25 variables that had strong relationship with the target variable were used for classification.

In this paper, we propose an extension of the work done by Chen et al. (2010). First, we replace the missing values in the financial data using the soft computing based data imputation approach proposed by Ankaiah and Ravi (2011). Then, we apply text mining on the textual (unstructured) data of phishing alerts. Thus, textual data is converted into structured data. Finally, we predict the risk level of phishing attacks using the combined financial data from the financial statement of the companies and textual data using MLP and DT separately. The work presented here is different from that of Chen et al. (2010) in that (i) we employ soft computing based imputation method for replacing the missing values in the financial variables unlike the mean substitution method used by them, and (ii) we employ PNN in addition to the DT and MLP methods used by them. We employ PNN because of its superlative performance that is reported by Ravisankar, Ravi, Raghava Rao, and Bose (2011) and Mohanty, Ravi, and Patra (2010).

The remainder of this paper is organized as follows: a brief review of literature on imputation of missing data is presented in Section 2. The methodology is explained in Section 3. Experimental setup is described in Section 4. Description of the dataset is presented in Section 5. Results and discussions are described in Section 6, followed by the conclusion in Section 7.

2. Review of data imputation techniques

Missing data handling methods can be broadly classified into two categories: deletion and imputation (Gheyas & Smith, 2010). The missing data ignoring techniques or deletion techniques simply delete the cases that contain missing data. Because of their simplicity, they are widely used and tend to be the default choice for most statistical packages, but this is not an effective solution. This approach has two forms: (i) listwise deletion that omits the cases or instances containing missing values. The main drawback of this method is that the application may lead to loss of large number of observations, which may result in high error and this is aggravated further if the original data set itself is too small (Song & Shepperd, 2007). (ii) Pairwise deletion method that considers each feature separately. For each feature, all recorded values are considered and missing data is ignored (Strike, El Emam, & Madhavji, 2001). This method is good when the overall sample size is small or missing data cases are large (Song & Shepperd, 2007).

Imputation method uses the available data to estimate the missing values. The simplest method of imputation is mean imputation, in which the missing values of a variable are replaced by the average value of all the remaining cases of that variable (Little & Rubin, 2002). The disadvantage of this method is that it ignores the correlations between various components (Schafer, 1997). When the variables are correlated, data imputation can be done using regression imputation. In regression imputation, regression equations are computed each time by considering the attribute containing incomplete value as the target variable. This method preserves the variance and covariance of missing data with other variables. Hot and cold deck imputation replaces the missing values with the closest complete components, where 'closest' is in terms of components that are present in both vectors for each case with a missing value (Schafer, 1997). The drawback of hot deck imputation is that the estimation of missing data is based on a single complete vector and thus it ignores the global properties of the dataset. The drawback of cold deck imputation is that missing values are replaced with the different dataset values (Little & Rubin, 2002). In the multiple imputation procedure, each missing value

is replaced by a set of reasonable and valid values, so that we get M complete data sets by replacing each value M times and by analyzing all datasets after which the combined inferences are made. According to Little and Rubin (2002), multiple imputation is better than case wise and mean substitution imputation. Regression methods are not as good as multiple imputations. Expectation maximization is an iterative process that continues until there is convergence in the estimation of parameters.

Another method that is used is the K-nearest neighbor (K-nn). In the K-nn approach the missing values are replaced by their nearest neighbors. The nearest neighbors are selected from the complete cases which minimize the distance function (Batista & Monard 2002; Batista & Monard, 2003; Jerez, Molina, Subirates, & Franco, 2006). Samad and Harp (1992) implemented the self-organizing approach for handling the missing data, which can be considered to be a variant of the k-nn approach.

Neural networks have been used for data imputation as well. In the neural network approach, MLP is trained as a regression model by using the complete cases and choosing one variable as target each time. By using an appropriate MLP model, each incomplete pattern is predicted. Several researchers have used MLP for missing data imputation (Gupta & Lam, 1996; Nordbotten, 1996; Sharpe & Solly, 1995; Yoon & Lee, 1999). Auto-associative neural network (AANN) is another machine learning technique that has been used for imputation. In AANN, the network is trained for predicting the inputs by taking the same input variables as target variables (Marseguerra & Zoia, 2002; Marwala & Chakraverty, 2006). Various imputation techniques that have been reported in extant literature are presented in Table 1.

3. Review of text mining

Text mining is a knowledge intensive process in which a user interacts with a collection of documents by using a suite of analytical tools. Text mining seeks to extract useful information from data sources through identification and exploration of interesting patterns (Srinivasan, 2003). Text mining algorithms operate on feature-based representations of the documents. Characters, words, terms, and concepts are the potential features used to represent the documents. Text mining gained popularity as a research tool due to its ability to mine the unstructured/digital content available on the internet. Text mining generally involves first converting the unstructured content into structured content before applying the usual data mining techniques. The free text of the phishing alerts is converted into structured data in the form of term-document matrix. A document is composed of several terms, the dimensionality of the document-term matrix will be very high if we use all the terms as attributes. So we group similar terms together so that the dimensionality of the document-term matrix is significantly reduced (Feldman & Sanger, 2007). Frequently occurring words that have similar meaning are grouped under a higher level concept. For example the terms 'credit union', 'agent bank', 'commercial bank' can be grouped under the concept 'bank'.

One of the important applications of text mining is in document management involving tasks such as text segmentation, key words extraction, indexing and text categorization. Clustering techniques and integrated information on personal preferences for document management was used by Wei, Chaing, and Wu (2006). A hybrid methodology that combined text mining with data mining has been adopted by some researchers. Text mining was used to analyze company news and discover social networks among companies (Ma, Sheng, & Pant, 2009). They utilized the discovered characteristics of the social networks to predict the revenue of the associated companies using decision trees and logistic regression. Document clustering has been used to identify the documents that showed disgruntled messages (Holton, 2009). He extracted key terms from the documents and used the terms as inputs to Bayesian networks to classify disgruntled and non-disgruntled messages. Text mining techniques such as key phrase extraction, cluster analysis, and concept links were used to discover different types of disclosures about information security incidents in financial reports.

4. Methodology

4.1. Soft computing approach for data imputation

The proposed missing data imputation approach is a two-stage approach. The block diagram (Fig. 1) depicts the hybrid imputation method. In this novel hybrid, K-means clustering is used for stage 1 (MacQueen, 1967). K-means is one of the simplest unsupervised learning algorithms that solve the clustering problem. The procedure for stage 1 imputation is as follows:

1. Identify K cluster centers by using the K-means clustering algorithm with complete instances.
2. Replace the incomplete instances by the nearest cluster center by measuring the distance with complete components of an incomplete record and the cluster centers.

The distance is measured by using the following formula:

$$d_j = \sum_{i=1}^{m} |x_i^{(j)} - C_j|^2$$

Table 1
Extant research related to techniques for data imputation.

Techniques with citation	Basic principle of operation
Listwise deletion Song and Shepperd (2007)	Eliminates all instances with missing values
Pairwise deletion Song and Shepperd (2007)	Eliminates instances only from those statistical analyses that require the information
Mean imputation Little and Rubin (2002))	Missing value is replaced by the mean
Regression imputation Little and Rubin (2002)	Estimates the relationships among the variables and then uses coefficients to estimate the missing values
Hot-deck imputation Schafer (1997)	Replaces the missing data with values from a similar complete data vector
Multiple imputation Little and Rubin (2002)	Replaces each missing value with a set of plausible ones that represent uncertainty about right value to impute
Expectation maximization Hsiao (1980)	An iterative procedure that continues until there is convergence in parameter estimates
Imputation with K-nearest neighbors Jerez et al. (2006)	K-nearest neighbors are selected from completed cases. The replacement value depends on the type of data: the mode can be used for discrete data and mean for continuous data
SOM imputation Samad and Harp (1992)	The value to be imputed is computed based on the activation group of nodes in the missing dimensions
MLP imputation Gupta and Lam (1996)	MLP is trained using only the complete cases as a regression model by taking incomplete variable as a target and remaining variables as input
Genetic algorithms and neural networks Marwala and Chakraverty (2006)	Genetic algorithm is used to minimize an error function derived from an auto associative neural network
Neuro-fuzzy neural networks Gabrys (2002)	Missing values are processed with general fuzzy min–max neural network architecture

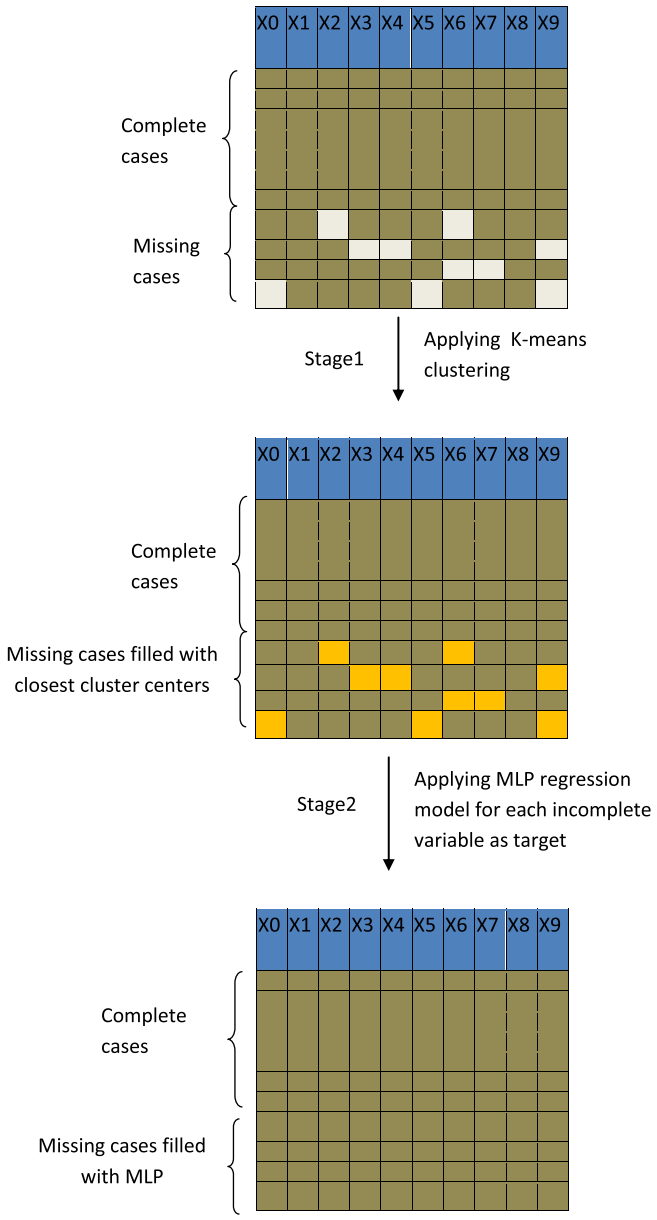

Fig. 1. Block diagram for the proposed two-stage data imputation technique (stage 1 uses K-means clustering and stage 2 uses MLP/PNN).

where j is the number of cluster centers, m is the number of complete components in each record (the value of m may change from one incomplete record to an other).

In the second stage, the approach proposed by Ankaiah and Ravi (2011) is used. MLP is a used for imputation. MLP is trained by using only complete cases. We train MLP as a regression model by taking one incomplete variable as the target variable and remaining variables as inputs. This procedure is repeated for as many times as the number of variables having missing values. Consequently, we train different MLP models that are equal to the number of incomplete variables in a given dataset. The steps for the MLP imputation scheme (stage 2) are as follows:

1. For a given incomplete dataset X, separate the instances that contain missing values from those without missing values. Take complete instances as variables with known values X_k and incomplete instances as variables with unknown values X_u.

Table 2
List of financial variables used in assessing severity of phishing alerts.

No.	Financial variable	No	Description of the financial variable
1	Expenditure in advertising	14	Value of tangible common equity
2	Book value per share	15	Total debt in current liabilities
3	Cost of goods sold	16	Number of employees
4	Earnings before interest and taxes	17	Income before extra-ordinary items
5	Total assets	18	Total amount of invested capital
6	Total inventories	19	Total long term debt
7	Total liabilities	20	Net loss in income
8	Total market value in fiscal year	21	Total operating expenses
9	Notes payable in short term borrowings	22	Total preferred/preference stock (capital)
10	Value of other intangibles	23	General sales and administrative expenses
11	Annual high price in fiscal year	24	Total revenue
12	Total receivables	25	Standard & Poor's core earnings
13	Total assets		

Table 3
List of terms extracted from textual description of phishing alerts.

No.	List of variables from text mining	No.	List of variables from text mining
1	Account	8	E-mail
2	Assets	9	Information
3	Bank	10	Person
4	Computer	11	Security
5	Confirmation	12	Update
6	Consumers	13	Warning
7	E-bay	14	Work

2. For each incomplete variable, train an MLP using X_k by considering remaining variables in X_k as inputs.
3. Predict the missing value, which is the target variable in the regression model of MLP, by using the initial approximate value which is given by K-means clustering in stage 1.

Repeat steps 2 and 3 for all incomplete variables.

4.2. Process for data imputation

For the financial dataset, the number of clusters, K, considered for imputation in stage-1 is fixed at 3, which is equal to number of risk levels for classifying phishing alerts. MLP employed in stage-2 has two parameters viz., learning rate (LR) and momentum rate (MR). Different imputed datasets are obtained for different combinations of LR and MR. In order to come out with an optimal imputed dataset, the strength of each of the alternatives (imputed datasets) is tested by invoking the classifiers DT and MLP using 10-fold cross validation. The imputed dataset that yields best classification accuracy is used for further experimentation.

4.3. Construction of document-term matrix for text mining

Key phrase extraction techniques are used to determine the important semantic concepts that can act as input variables for classification. The file containing the textual data for phishing alerts is taken into a text document and the letters in the entire document are converted to lower case. The text document is fed to the data mining tool Rapidminer to identify the important terms in the textual data for phishing. The frequency of occurrence of each term in the entire text document is calculated. The terms that

Table 4
Confusion matrix for financial data alone.

	Our results						Chen et al. (2010) results					
	MLP			DT			MLP			DT		
	High	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium
High	**48 (55.17%)**	11	28	**67 (77.01%)**	8	12	47 (54.02%)	16	23	59 (67.81%)	11	16
Low	0	20 (86.95%)	3	0	20 (86.95%)	3	0	**21 (91.3%)**	2	0	**22 (95.65%)**	1
Medium	236	140	**542 (59.04%)**	285	120	513 (55.88%)	214	212	495 (53.92%)	218	126	**577 (62.85%)**
Overall accuracy	**610 (59.33%)**			600 (58.36%)			563 (54.76%)			**658 (64%)**		

Table 5
Confusion matrix for textual data alone.

	Our results						Chen et al. (2010) results					
	MLP			DT			MLP			DT		
	High	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium
High	**69 (79.31%)**	6	12	**71 (81.6%)**	5	11	55 (63.21%)	8	23	64 (73.56%)	4	18
Low	1	**21 (91.3%)**	1	1	**21 (91.3%)**	1	2	19 (82.6%)	2	3	18 (78.26%)	2
Medium	317	42	559 (60.89%)	268	28	**622 (67.75%)**	171	101	**649 (70.69%)**	234	82	605 (65.9%)
Overall accuracy	649 (63.3%)			**714 (69.4%)**			**723 (70.37%)**			687 (66.82%)		

Table 6
Confusion matrix for combined financial and textual data.

	Our results						Chen et al. (2010) results					
	MLP			DT			MLP			DT		
	High	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium
High	**81 (93.1%)**	3	3	**84 (95.65%)**	2	1	65 (74.71%)	3	18	75 (86.2%)	5	6
Low	1	**22 (95.65%)**	0	0	22 (95.62%)	1	1	18 (78.26%)	4	0	**23 (100%)**	0
Medium	145	35	**738 (80.39%)**	165	14	**739 (80.5%)**	181	34	706 (76.9%)	133	64	724 (78.86%)
Overall accuracy	**841 (81.80%)**			**845 (82.19%)**			789 (76.60%)			822 (79.81%)		

Table 7
Confusion matrix for financial, textual, and combined data using the classifier PNN.

	Financial data			Textual data			Combined data		
	High	Low	Medium	High	Low	Medium	High	Low	Medium
High	65 (74.12%)	9	13	47 (50.02%)	7	32	70 (80.45%)	6	11
Low	0	20 (86.95%)	3	1	21 (91.30%)	1	0	22 (95.65%)	1
Medium	292	105	521 (56.75%)	192	50	**676 (73.63%)**	66	95	**757 (82.46%)**
Overall accuracy	606 (58.94%)			**744 (72.3%)**			**849 (82.58%)**		

are semantically important and have relatively high frequency than others are considered to be the important ones. Some of the frequently occurring words have almost similar meaning. So we grouped such concepts together under a higher level concept. For example the words 'client', 'member', 'customer' are grouped under the concept 'customer'. We used 14 terms (see Table 3) for constructing the document-term matrix. This procedure is similar to that used by Chen et al. (2010).

A document-term matrix or a term-document matrix is a matrix that has entries as the frequency of terms, which occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to the terms. In the phishing dataset, there are 1028 instances for constructing the document-term matrix. We considered each record as a document and calculated the frequency of occurrence of 14 concepts within each document. For constructing the document

term matrix, each record is read and the record is converted to meaningful text by eliminating the stop words and delimiters. The meaningful text thus obtained is split into words and the words are compared with the 14 concepts obtained above. The frequency of occurrence of each concept is calculated and is tabulated as a document-term matrix. By performing this analysis the natural language of phishing alert is converted to structured data that can be used as input to classification models. We combined the variables from the financial data and the variables from the document-term matrix to form a dataset with 40 variables. The resulting dataset is used for the task of classification.

5. Experimental design

Chen et al. (2010) chose top 25 financial variables for the classification purpose (see Table 2). We also followed the same approach. The total number of instances is 1028. It is a three class classification problem. All the missing values are imputed by using the 2-stage approach. The dependent variable in our method is the measure of severity of phishing alerts i.e., the risk level of the phishing alert. The severity of phishing attacks are classified into three risk levels: high, medium and low. MLP and DT are employed due to their reported superior performance in other applications related to information security and to make the comparison of our results with that of Chen et al. (2010) more meaningful. The risk levels for the phishing alerts are not evenly distributed. Therefore, we oversampled the high risk and low risk instances of data but kept the medium risk instances the same so that the distribution of the three groups became 1:1:1 in the dataset. Thus, all the three risk groups had 918 instances in the modified dataset. For building the classification model, 70% of the oversampled data is used for training and 30% is used for testing. However, for validation purpose, we used the original data in its unbalanced form as it is. We repeated this experiment 10 times and in each iteration we changed the composition of the training and test sets randomly and calculated the average accuracy of the model over these 10 fold cross-validation.

6. Results and discussion

We used MLP, DT, and PNN as classification models. We compared our results with that of Chen et al. (2010). We used an open source data mining tool KNIME (www.knime.org) for implementing MLP, DT, and PNN. We used another open source data mining tool Rapidminer (www.rapid-I.com) for performing the data preparation phase.

The average results of 10 folds using MLP, DT, and PNN in the case of financial, textual and combined data are presented in Tables 3–7. In each Table, the best results are represented with a bold font. Chen et al. (2010) reported overall accuracy of 79.81% and 76.60% for the combined textual and financial data using DT and MLP respectively (see Table 6). However, we obtained overall accuracies of 82.58%, 82.19%. and 81.80% for PNN, MLP, and DT respectively (see Tables 6 and 7). Thus, the results of the present study outperformed that of Chen et al. (2010).

For financial data alone, the accuracies reported by Chen et al. (2010) using MLP are 54.02%, 91.30%, and 53.92% for the risk levels high, low, and medium respectively. However, we achieved an accuracy of 55.17%, 86.95%, and 59.04% for the three risk levels using MLP (see Table 4). For textual data alone, our results of 79.31%, and 91.30% turned out to be superior to 63.21% and 82.60% reported by Chen et al. (2010) for risk levels high and low respectively. However, for medium risk level, we obtained an accuracy of 60.89% which is lower than 70.69% reported by Chen et al. (2010) (see Table 5).

Using DT, for financial data alone, we obtained a classification accuracy of 77.01% which is superior to 67.81% reported by Chen et al. (2010) for high risk level. However, our results of 86.95% and 55.85% are inferior to 95.65% and 62.85% reported by Chen et al. (2010) for the risk levels low and medium respectively (see Table 4). In addition, when considering textual data alone and using DT, the accuracies of 81.6%, 91.3% and 67.75% reported in this paper are superior to 73.56%, 78.26%, and 65.90% reported by Chen et al. (2010) for the risk levels high, low, and medium respectively (see Table 5).

Using MLP, the overall accuracies of 93.10%, 95.65% and 80.39% that we obtained are much superior to 74.71%, 78.26%, and 76.90% reported by Chen et al. (2010) for the risk levels high, low, and medium respectively (see Table 6). Using DT, Chen et al. (2010) reported classification accuracies of 86.20% and 78.86% for the risk levels high and medium respectively, whereas we obtained an accuracy of 95.65% and 80.5% which are superior to Chen et al. (2010). However, for low risk level we achieved slightly lower accuracy of 95.62% (22 out of 23) compared to 100% (23 out of 23) reported by Chen et al. (2010). The number of instances with low risk level in the dataset is very low compared to that of instances with high and medium risk level instances and this may have influenced the classification accuracies for low risk in this research.

The classification results using PNN are shown in Table 7. It turns out that the overall accuracy of 82.58% for the combined data using PNN is better than MLP and DT results reported by us and by Chen et al. (2010). The superiority of PNN as a classifier can also be seen in the case of textual data alone, where the overall accuracy is 72.3% and beats the accuracies obtained by MLP and DT in this paper and also those reported by Chen et al. (2010). However, for financial data alone the overall accuracy using PNN is not the best. For the medium risk level, the overall accuracy obtained using combined data (i.e., 82.46%) and textual data alone (73.63%) superseded the overall accuracies obtained using MLP and DT in our research and those reported by Chen et al. (2010). These observations make us believe that it is advisable to explore the use of PNN in addition to MLP and DT for classification of risk level of phishing alerts.

7. Conclusion

We assessed the severity of phishing alerts in financial companies by resorting to soft computing based data imputation, and hybrid data and text mining. The classification accuracy using both textual data and the imputed financial data is 82.19% using DT, and 81.80% using MLP. The results in both cases turned out to be superior to those of Chen et al. (2010). We concluded that imputing the missing values in financial data using the method proposed by Ankaiah and Ravi (2011) yielded better results than mean imputation followed by Chen et al. (2010). Another important conclusion is that the present method yielded significantly higher accuracies in high and medium risk level alerts when compared to that of Chen et al. (2010). Further, we demonstrated the need to consider a classifier like PNN for this research context. PNN employed in this study as a classifier outperformed both DT and MLP in terms of overall accuracy for classifying the risk level of phishing alerts.

Appendix A

A.1. Overview of PNN

PNN is a feed-forward neural network involving a one pass training algorithm used for classification and mapping of data. PNN was introduced by Specht (1990). PNN is an implementation of the statistical algorithm called kernel discriminant analysis in

which the operations are organized into multilayer feed forward network with four layers: input layer, pattern layer, summation layer, and output layer.

It is a pattern classification network based on the classical Bayes classifier, which is statistically an optimal classifier that seeks to minimize the risk of misclassifications. Any pattern classifier places each observed data vector $x = [x_1, x_2, x_3, \ldots, x_N]^T$, into one of the predefined classes c_i, $i = 1, 2, \ldots, m$ where m is the number of possible classes. The effectiveness of any classifier is limited by the number of data elements that the vector x can have and the number of possible classes m. The classical Bayes pattern classifier implements the Bayes conditional probability rule that the probability $P(ci|x)$ of x being in class c_i is given by:

$$P(c_i/\mathbf{x}) = \frac{P(\mathbf{x}/c_i)P(c_i)}{\sum_{j=1}^{m} P(\mathbf{x}/c_j)P(c_j)}$$

where $P(x|c_i)$ is the conditional probability density function of x given set c_i, $P(c_j)$ is the probability of drawing data from class c_j. Vector x is said to belong to a particular class c_i, if $P(c_i|x) > P(c_j|x)$, $\forall j = 1, 2, \ldots, m$ and j is not equal to i. This input x is fed into each of the patterns in the pattern layer. The summation layer computes the probability $P(c_i|x)$ of the given input x to be in each of the classes c_i that is represented by the patterns in the pattern layer. The output layer selects the class for which the highest probability is obtained in the summation layer. The input is then made to belong to this class.

The effectiveness of the network in classifying input vectors depends highly on the value of the smoothing parameter σ. PNN is employed in addition to DT and MLP because of the fact that it can train fast on sparse data sets, and it is a universal approximator for smooth classification problems (Ravisankar et al., 2011).

References

Abdella, M., & Marwala, D. (2005). The use of genetic algorithms and neural networks to approximate missing data in database. In *Proceedings of the IEEE 3rd international conference on computational cybernetics (ICCC)* (pp. 207–212).

Airoldi, E., & Malin, B. (2004). Data mining challenges for electronic safety: The case of fraudulent intent detection in e-mails. In *Proceedings of the workshop on privacy and security aspects of data mining* (pp. 57–66).

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

APWG (2009). Phishing activity trends report second half 2008, Anti-phishing working group, pp. 1–12.

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

Batista, G., & Monard, M. C. (2003). Experimental comparison of K-nearest neighbor and mean or mode imputation methods with the internal strategies used by C4.5 and CN2 to treat missing data. Technical Report, University of Sao Paulo.

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.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge, UK: Cambridge University Press.

Gabrys, B. (2002). Neuro-fuzzy approach to processing inputs with missing values in pattern recognition problems. *International Journal of Approximate Reasoning, 30*, 149–179.

Garcıa-Laencina, P. J., Sancho-Gomez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: A review. *Neural Computing & Applications, 19*, 263–282.

Gheyas, I. A., & Smith, L. S. (2010). A neural network-based framework for the reconstruction of incomplete data sets. *Neuro Computing, 73*(16), 3039–3065.

Gupta, A., & Lam, M. S. (1996). Estimating missing values using neural networks. *Journal of the Operational Research Society, 47*(2), 229–238.

Hai, W., & Shouhong, W. (2010). The use of ontology for data mining with incomplete data. *Principle Advancements in Database Management Technologies*, 375–388.

Holton, C. (2009). Identifying disgruntled employee systems fraud through text mining: A simple solution for multi-billion dollar problem. *Decision Support Systems, 46*(4), 853–864.

Hsiao, C. (1980). Missing data and maximum likelihood estimation. *Economic Letters, 6*, 249–253.

Jackobsson, M., & Myers, S. (2007). *Phishing and countermeasures: Understanding the increasing problem of electronic identity theft*. Hoboken, NJ, USA: Wiley-Interscience.

Jagatic, T., Johnson, N., & Menczer, F. (2006). Social phishing. *Communications of the ACM, 50*(10), 1–10.

Jerez, J., Molina, I., Subirates, J., & Franco, L. (2006). Missing data imputation in breast cancer prognosis. In *Proceedings of the 24th IASTED international conference on biomedical engineering (BioMed'06)*, Anaheim, CA, USA.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ, USA: Wiley-Interscience.

Ma, Z., Sheng, O. R. L., & Pant, G. (2009). Discovering company revenue relations from news: A network approach. *Decision Support Systems, 47*(4), 408–414.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, USA, (pp. 281–297).

Marseguerra, M., & Zoia, A. (2002). The auto-associative neural network in signal analysis II. Application to on-line monitoring of a simulated BWR component. *Annals of Nuclear Energy, 32*(11), 1207–1223.

Marwala, T., & Chakraverty, S. (2006). Fault classification in structures with incomplete measured data using auto associative neural networks and genetic algorithm. *Current Science India, 90*(4), 542–548.

Merlin, P., Sorjamaa, A., Maillet, B., & Lendasse, A. (2010). X-SOM and L-SOM: A double classification approach for missing value imputation. *Neurocomputing, 73*, 1103–1108.

Mohanty, R., Ravi, V., & Patra, M. R. (2010). Web services classification using intelligent techniques. *Expert Systems with Applications, 37*(7), 5484–5490.

Nordbotten, S. (1996). Neural network imputation applied to the Norwegian 1990 population census data. *Journal of Official Statistics, 12*, 385–401.

Ravisankar, P., Ravi, V., Raghava Rao, G., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems, 50*, 491–500.

Samad, T., & Harp, S. A. (1992). Self-organization with partial data network. *Computation in Neural Systems, 3*, 205–212.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Florida, USA: Chapman & Hall.

Sharpe, P. K., & Solly, R. J. (1995). Dealing with missing values in neural network based diagnostic systems. *Neural Computing & Applications, 3*(2), 73–77.

Singh, N. P. (2007). Online frauds in banks with phishing. *Journal of Internet Banking and Commerce, 12*(2), 1–27.

Song, Q., & Shepperd, M. (2007). A new imputation method for small software project data sets. *Journal of Systems and Software, 80*(1), 51–62.

Specht, D. A. (1990). Probabilistic neural networks. *Neural Networks, 3*(10), 109–118.

Srinivasan, P. (2003). Text mining: Generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology, 55*(5), 396–413.

Strike, K., El Emam, K., & Madhavji, N. (2001). Software cost estimation with incomplete data. *IEEE Transactions on Software Engineering, 27*(10), 890–908.

Wei, C. P., Chaing, R. H. I., & Wu, C. C. (2006). Accommodating individual preferences in the categorization of documents: A personalized approach. *Journal of Management Information Systems, 23*(2), 173–201.

Workman, M. (2008). Wisecarckers: A theory-grounded investigation of phishing and pretext social engineering threats to information security. *Journal of the American Society for Information Science and Technology, 59*(4), 662–674.

Yoon, S. Y., & Lee, S. Y. (1999). Training algorithm with incomplete data for feed-forward neural networks. *Neural Processing Letters, 10*, 171–179.