Measuring the Impact of Imputation in Financial Fraud

Stephen O. Moepya[1,2(✉)], Sharat S. Akhoury[1], Fulufhelo V. Nelwamondo[1,2], and Bhekisipho Twala[2]

[1] CSIR Modeling and Digital Science, Pretoria, South Africa
{smoepya,sakoury,FNelwamondo}@csir.co.za
http://www.csir.co.za

[2] Faculty of Engineering and the Built Environment, University of Johannesburg, Johannesburg, South Africa
btwala@uj.ac.za

Abstract. In recent years, data mining techniques have been used to identify different types of financial frauds. In some cases, the fraud domain of interest contains data with missing values. In financial statement fraud detection, instances which contain missing values are usually discarded from the analysis. This may lead to crucial information loss. Imputation is a technique to estimate missing values and is an alternative to case-wise deletion. In this paper, a study on the effectiveness of imputation is taken using financial statement fraud data. Also, the measure of similarity to the ground truth is examined using five distance metrics.

Keywords: Financial statement fraud · Missing values · Imputation · Distance metrics

1 Introduction

Financial statement fraud (FSF) is a deliberate and wrongful act carried out by public or private companies using materially misleading financial statements that may cause monetary damage to investors, creditors and the economy. The two most prominent examples of financial statement fraud are Enron Broadband and Worldcom. In these two cases, wronged investors and the financial market suffered as a result of this type of fraud. The collapse of Enron alone caused a $70 billion market capitalization loss. The Worldcom scandal, caused by alleged FSF, is the biggest bankruptcy in United States history [13]. In this light, the application domain of financial statement fraud detection has attracted a keen interest.

Public companies are required to issue audited financial statements at least once a year. This requirement is to allow for *standardized* comparability between companies. Financial statements contain information about the financial position, performance and cash flows of a company [10]. The statements also inform

M. Núñez et al. (Eds.): ICCCI 2015, Part II, LNCS 9330, pp. 533–543, 2015.
DOI: 10.1007/978-3-319-24306-1_52

readers about related party transactions. In practice, financial statement fraud (FSF) might involve [16]: the manipulation of financial records; intentional omission of events, transactions, accounts, or other significant information from which financial statements are prepared; or misapplication of accounting principles, policies, and procedures used to measure, recognize, report, and disclose business transactions.

Missing data, in general, is a common feature in many practical applications. In some cases, the data in FSF domain may contain missing values. Two approaches are taken when encountered with missing values: exclusion of all instances with missing values (casewise deletion) or an estimation of values for all missing items (imputation). Data can be missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR) [14].

In this paper, the objective is to investigate the use of imputation techniques using authentic financial fraud data. Seven imputation techniques are investigated in order to test the significance of imputation. Imputed values will be evaluated via distance metric scores to attain a measure of similarity to known values ('ground truth'). The remainder of this paper is structured as follows. Section 2 reviews related research. Section 3 provides a brief description of the methodology. Sections 4 and 5 describe the sample data, experimental setup and details the results thereof. Finally, Section 6 presents the conclusion.

2 Related Work

In this section, a brief description of the impact of imputation in the finance domain follows.

Sorjamaa et al. [15] present a combination of Self-Organizing Maps (SOM) to treat missing values using corporate finance data. The study used data involving approximately 6000 companies listed on either the Paris or London Stock Exchange during the period 1999-2006. The authors concluded that using a combination of SOMs (instead of the traditional SOM imputation) yields better performance based on test error. Additional imputation schemes were not considered in this study to benchmark performance.

The use of multiple imputation (MI) on the consumer's choice between debit and credit cards is presented by King [9]. The dataset used in the experiment was from the 1998 Survey of Consumer Finances. The results suggest that the persistence of debit cards is due to the fact that even households that use credit cards without borrowing do not view credit as a substitute for debit. A limitation of the experiment was using only 5 imputed datasets for MI. Increasing the number of datasets used could have lead to a reduction in bias.

Fogarty [5] analyzed the importance of utilizing imputation to enhance credit score cards. The focus of this study was treating reject inference as a missing data problem. The data used was collected from a large German consumer finance company and includes 'accepts' and 'rejects' from the businesses application data and an associated credit performance variable from the behavioral data. The author concluded that model-based MI is an enhancement over traditional

missing data approaches to reject inference. The effect of imputation was not tested against a benchmark dataset.

An interesting imputation approach on credit scoring data is presented by Paleologo et al. [11]. The authors' aim was to build and validate robust credit models on Italian company credit request data. Missing values were replaced with either the minimum or maximum value from the corresponding feature. There was no attempt to utilize standard imputation techniques to check predictive performance.

In light of the literature reviewed, the importance of imputing financial data has been highlighted. Imputation seems to be a viable option instead of case-wise deletion. Hence in this study, an investigation of several imputation methods using financial statement fraud data is undertaken. The similarity between a ground truth and imputed data will be computed using distance metrics.

3 Methodology

3.1 Imputation

The imputation of missing values can be broadly split into two categories: statistical and machine learning based imputation. Statistical imputation includes methods such as mean imputation, hot-deck and multiple imputation methods based on regression and the expectation maximization (EM) algorithm [6]. Machine learning approaches for imputing missing values create a predictive model to estimate missing values. These methods model the missing data estimation based on the available information in the data. For example, if the observed dataset contains some useful information for predicting the missing values, the imputation procedure can utilize this information and maintain a high precision. This section gives a description of both statistical and machine learning based imputation methods.

Mean imputation is one of the simplest methods to estimate missing values. Consider a matrix X containing a full data set. Suppose that the value x_{ij}belongs to the kth class C_k and it is missing. Mean imputation replaces x_{ij} with $\bar{x}_{ij} = \sum_{i:x_{ij} \in C_k} \frac{x_{ij}}{n_k}$, where n_k represents the number of non-missing values in the jth feature of the kth class.

In kNN imputation [8], missing cases are imputed using values calculated from corresponding k-nearest neighbors. The nearest neighbor of an arbitrary missing value is calculated by minimizing a distance function. The most commonly used distance function is the Euclidean distance between two instances y and z as $d(y,z) = \sqrt{\sum_{i \in D}(x_{yi} - x_{zi})^2}$, where D is a subset of the matrix X containing all instances without any missing values. Once k-nearest neighbors are computed, the mean (or mode) of the neighbors is imputed to replace the missing value.

Principal Component Analysis (PCA) imputation involves replacing missing values with estimates based on a PCA model. Suppose that the columns of matrix X are denoted by d-dimensional vectors $y_1, y_2, \cdots, y_n$. PCA imputation assumes that these vectors can be modeled as $y_j \approx Wz_j + m$, where W is a

$d \times c$ matrix, z_j are the c-dimensional vectors of principal components and m is a bias vector. This imputation method iterates and converges to a threshold by minimizing the error $C = \sum_{j=1}^{n} \|y_j - Wz_j - m\|^2$.

The Expectation-Maximization (EM) is an iterative procedure that computes the maximum likelihood estimator (MLE) when only a subset of the data is available. Let $X = (X_1, X_2, \cdots, X_n)$ be a sample with conditional density $f_{x|\Theta}(x|\theta)$ given $\Theta = \theta$. Assume that X has missing variables $Z_1, Z_2, \cdots, Z_{n-k}$ and observed variables $Y_1, Y_2, \cdots, Y_k$. The log-likelihood of the observed data Y is

$$l_{obs}(\theta; Y) = \log \int f_{X|\Theta}(Y, z|\theta) v_z(dz). \tag{1}$$

To maximize l_{obs} with respect to θ, the E-step and M-step routines are used. The E-step finds the conditional expectation of the missing values given observed values and current estimates of parameters. The second step, the M step, consists of finding maximum likelihood parameters as though the missing values were filled in [12]. The procedure iterates until convergence.

Singular Value Thresholding (SVT) [3] is a technique that has been used in Exact Matrix Completion (MC). The SVT algorithm solves the following problem:

$$\min_{X \in C} \tau \|X\|_* + \frac{1}{2}\|X\|_F^2, s.t.\ \mathcal{A}_{\mathcal{I}}(X) = \mathcal{A}_{\mathcal{I}}(M), \tag{2}$$

where $\tau \geq 0$ and the first and second norms are the nuclear and Frobenius norms respectively. M is an approximately low rank matrix. In the above equation, $\mathcal{A}$ is the standard matrix completion linear map where $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^k$. SVT is comprised of following two iterative steps:

$$\begin{cases} X_t = \mathcal{D}_t(A^*_{\mathcal{I}}(y_{t-1})) \\ y_t = y_{t-1} - \delta(A_{\mathcal{I}}(X_t) - b). \end{cases} \tag{3}$$

In the above equation, the shrinkage operator $\mathcal{D}_\tau$, also know as the soft-thresholding operator, is denoted as $\mathcal{D}_\tau = U\Sigma_\tau V^T$ where U and V are matrices with orthonormal columns and $\Sigma_\tau = diag(\max\{\sigma_i - \tau, 0\})$ with $\{\sigma_i\}_{i=1}^{\min\{n_1, n_2\}}$ corresponding to the singular values of the decomposed matrix. The step size of the iterative algorithmic process is given by δ.

Random Forests (RF), introduced by Breiman [1], is an extension of a machine learning technique named bagging which uses Classification and Regression Trees (CART) to classify data samples. Imputation via RF begins by imputing predictor means in place of the missing values. A RF is subsequently built on the data using roughly imputed values (numeric missing values are re-imputed as the weighted average of the non-missing values in that column). This process is repeated several times and the average of the re-imputed values is selected as the final imputation.

A brief explanation of Singular Value Decomposition (SVD) imputation follows. Consider the SVD of a matrix $X \in \mathbb{R}^{n_1 \times n_2}$ of rank r. In this instance, $X = U\Sigma V$, U and V are $n_1 \times r$ and $n_2 \times r$ orthogonal matrices respectively and $\Sigma = diag(\{\sigma_i\}_{1 \leq i \leq r})$. The σ_is are known as the positive singular values.

SVD imputation begins by replacing all missing values with some suited value (mean or random). The SVD is computed and missing values replaced with their prediction according to SVD decomposition. The process is repeated until the imputed missing data fall below some threshold.

3.2 Distance Metrics

In a formal sense, distance can be defined as follows. A distance is a function d with non-negative real values, defined on the Cartesian product $X \times X$ of a set X. It is termed a metric on X if $\forall\, x, y, z \in X$ it has the following properties:

1. $d(x, y) = 0 \iff x = y$;
2. $d(x, y) + d(y, z) \geq d(x, z)$; and
3. $d(x, y) = d(y, x)$.

Property 1 asserts that if two points x and y have a zero distance then they must be identical. The second property is the well known triangle inequality and states, given three distinct points x, y and z, the sum of two sides xy and yz will always be greater than or equal to side xz. The last property states that the distance measure is symmetrical.

Table 1. Distance/Similarity Metrics

Lorentzian	$d(x,y) = \sum_{i=1}^{n} \ln(1 + \lvert x_i - y_i \rvert)$
Minowski	$d(x,y) = \sqrt[p]{\sum_{i=1}^{n} \lvert x_i - y_i \rvert^p}$
Dice	$d(x,y) = \frac{\sum_{i=1}^{n} (x_i - y_i)^2}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2}$
Squared Euclidean	$d(x,y) = \sum_{i=1}^{n} (x_i - y_i)^2$
Motyka	$d(x,y) = \frac{\sum_{i=1}^{n} \max(x_i, y_i)}{\sum_{i=1}^{n} (x_i + y_i)}$

Table 1 presents the definition of five distance metrics used in this experiment. The above similarity metrics each represent five of the eight types of similarity families. The data that will be used is not suitable for the Squared-chord, Shannon's entropy and combination families since it contains negative real values. The following section presents the results of the experimentation.

4 Data Description and Experimental Setup

4.1 Data

The dataset used in this experiment was obtained from INET BFA, one of the leading providers of financial data in South Africa. The data comprises publicly

Table 2. Summary statistics of the selected variables in the data

Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Assets-to-Capital Employed	-26.120	1.060	1.240	1.612	1.60	224.30
Book Value per Share	-966810	79	418	22279	1610	49596137
Cash flow per Share	-7198	8	84	3177	396	5159700
Current Ratio	0.010	1.00	1.410	3.283	2.251	726.130
Debt to Assets	0.0	0.270	0.4800	0.8804	0.680	1103.00
Debt to Equity	-182.370	0.350	0.820	2.516	1.730	760.940
Earnings per Share	-460338.5	2.3	44.5	391.6	218.2	825937.7
Inflation adjusted Profit per Share	-9232	1	43	2858	239	4898104
Inflation adjusted Return on Equity	-87472.97	3.11	13.45	-55.60	23.37	17063.16
Net Asset Value per Share	-373040	60	405	29438	1817	66658914
Quick Ratio	0.01	0.730	1.050	2.949	1.720	726.130
Retention Rate	-7204.35	57.67	89.39	70.97	100.00	5214.29
Return on Equity	-13600.00	4.045	14.830	-3.549	25.420	17063.160
Return on Capital Employed	-13600.00	1.500	8.700	-0.551	17.415	6767.330

listed companies on the Johannesburg Stock Exchange (JSE) between years 2003 and 2013. The different sectors for the listed companies on the JSE are: Basic Materials; Consumer Goods; Consumer Services; Financial; Health Care; Industrial; Oil and Gas; Technology; Telecommunications and Utilities. In the data, 123 (out of 3043) companies were known to have received a qualified financial report by an accredited auditor.

Each of the variables in Table 2 represent an aspect of measuring company performance. For example, the 'Quick Ratio' captures the amount of liquid assets per unit current liability. This ratio is a measure of how quickly a company can pay back its short-term debt. 'Earnings per share' (EPS) is a common metric for company valuation. It is a representation of a company's earnings, net of taxes and preferred stock dividends, that is allocated to each common stock. The other ratios fall into either the *profitability*, *solvency* or *leverage* category.

4.2 Experimental Setup

The experiments for this paper were conducted on a Intel(R) Core (TM) i5-3337U CPU @ 1.80 GHz with 6 GB memory. The implementation for algorithms are performed using the following R packages: 'imputation'[1] 'Amelia' [7], 'randomForest' [2] and 'yaImpute' [4].

The impact of imputation will be tested by running a Monte Carlo (MC) simulation (with 100 trails) as follows. In every MC trail:

1. Create 6 levels of missingness randomly from the 'ground truth' dataset using 1%, 2%, 5%, 10%, 15% and 20% as missing proportion;
2. Impute missing values on each missingness level using SVD, kNN, PCA, SVT, Mean, EM and RF imputation;
3. Compute the distance between the imputed values and the corresponding 'ground truth' values using the 5 distance/similarity measures (given in previous Section)

[1] This package has been now been archived in CRAN repository.

Once the Monte Carlo simulation is complete, compute the mean and standard deviations of the distances (from the 'ground truth') to evaluate the performance of the imputation techniques. The distances are evaluated using the five distance metrics.

The imputation schemes have parameters that needed to be selected. It was decided not to stray too far from the default parameter settings for the imputation methods. For PCA imputation: max iterations was set to 1000, number of principal components was set to 2 and a $1e-06$ threshold. The rank approximation, r, was set the value 3 for SVD imputation. The Random Forest default parameters are as follows: $ntree = 300$ and $iter = 5$, these are the number of trees and iterations respectively. The default tolerance value for the EM imputation is $1e-04$, and the value $k = 3$ was selected for kNN imputation.

5 Results

In this section, the results of the Monte Carlo simulation are presented. An analysis of the performance of the distance metrics will be undertaken.

Normalized distances of the 7 imputation methods using Lorentzian and Squared Euclidean distance

Table 3. Lorentzian

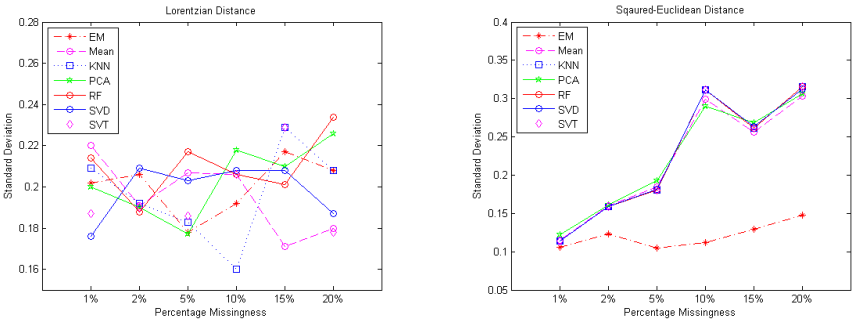
Method	1%	2%	5%	10%	15%	20%
EM	0.523	0.548	0.667	0.524	**0.487**	0.448
kNN	0.453	0.400	0.494	0.438	0.520	**0.395**
Mean	0.412	0.418	0.545	0.648	0.594	0.673
PCA	0.409	0.382	0.345	0.441	0.493	0.509
RF	0.486	0.399	0.437	0.491	0.560	0.533
SVD	**0.256**	**0.345**	**0.294**	**0.405**	0.504	0.536
SVT	0.445	0.412	0.457	0.492	0.550	0.412

Table 4. Squared Euclidean

Method	1%	2%	5%	10%	15%	20%
EM	0.018	**0.023**	**0.015**	**0.037**	**0.043**	**0.049**
kNN	0.019	0.035	0.067	0.168	0.191	0.277
Mean	**0.017**	0.032	0.050	0.147	0.165	0.239
PCA	0.025	0.039	0.069	0.165	0.201	0.285
RF	**0.017**	0.032	0.050	0.147	0.165	0.239
SVD	**0.017**	0.032	0.050	0.147	0.166	0.236
SVT	**0.017**	0.032	0.050	0.147	0.165	0.239

Table 3 presents the Lorentzian distance for each of the seven imputation techniques. Each column in the table represents the six levels of missingness that was randomly generated during the Monte Carlo simulation. Smaller values represent closer similarity to the ground truth. In this case, it can be seen that SVD imputation has the lowest normalized average distance for missingness levels $\leq 10\%$. For missingness above 10%, EM imputation and kNN produce the lowest average normalized distance with a score of 0.487 and 0.395 respectively. For all levels of missingness, Figure 1a shows a different imputation scheme attains the lowest standard deviation over the 100 Monte Carlo trails for the Lorentzia distance metric.

The Squared Euclidean distance results are presented in Table 4. At the 1% level of missingness, there is a four-way tie between the Mean, RF, SVD and SVT each achieving an average normalized distance of 0.17. For missingness greater than 1%, EM produced the lowest average distance values. These values seem to be substantially lower than distances of other imputation techniques, i.e., at the 20% missingness level EM achieves 0.049 and all other schemes have values



(a) Lorentzian **(b)** Squared Euclidean

Fig. 1. Normalized standard deviation distances of the 7 imputation methods using Lorentzian and Squared Euclidean distance

greater than 0.2. Figure 1b shows that the minimum average standard deviation, for all levels of missingness, is produced by EM.

Normalized distances of the 7 imputation methods using Dice and Manhattan distance

Table 5. Dice

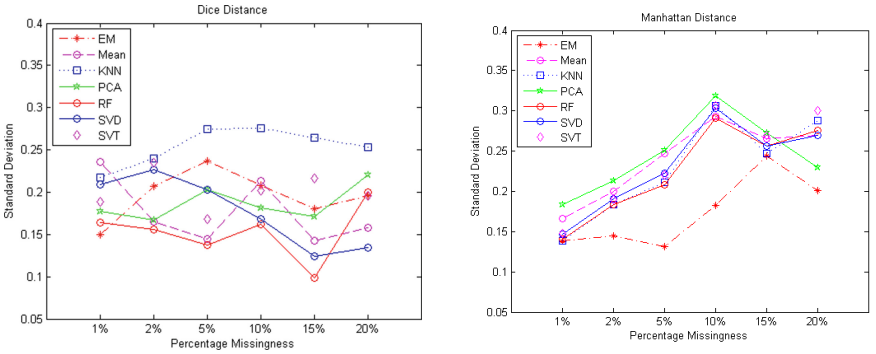
Method	1%	2%	5%	10%	15%	20%
EM	0.526	0.577	0.680	0.541	0.384	0.581
kNN	0.820	0.902	0.908	0.906	0.876	0.940
Mean	0.376	0.508	0.642	0.754	0.801	0.840
PCA	0.312	**0.297**	0.391	**0.391**	0.466	0.519
RF	**0.265**	0.467	**0.249**	0.526	**0.259**	0.834
SVD	0.297	0.352	0.420	0.440	0.417	**0.448**
SVT	0.821	0.766	0.835	0.792	0.649	0.864

Table 6. Manhattan

Method	1%	2%	5%	10%	15%	20%
EM	0.111	0.168	**0.121**	**0.201**	0.284	**0.212**
kNN	0.093	0.130	0.216	0.359	0.370	0.483
Mean	**0.065**	**0.085**	0.128	0.241	**0.248**	0.316
PCA	0.084	0.122	0.186	0.405	0.454	0.446
RF	0.072	0.100	0.163	0.298	0.275	0.390
SVD	0.073	0.108	0.170	0.325	0.346	0.418
SVT	0.068	0.101	0.153	0.283	0.282	0.365

Table 5 shows that RF achieves the lowest normalized average distance for missingness levels: 1%, 5% and 15%. PCA outperforms all other schemes for missingness level 2% and 10%. SVD attains the lowest value for the highest missingness level. At 20% missingness, EM, SVD and PCA attain average distances of lower than 0.6 whilst the other schemes score above 0.8. RF produces the lowest average standard deviation for most of the missingness levels. This is seen in Figure 2a where it achieves the least amount of variation in four out of the six levels using the dice distance metric.

The results for Manhattan distance are given in Table 6. These result bare some similarity to Table 5, where the similarity measure favors two imputation schemes. EM and Mean imputation achieve the lowest distance values levels 1%, 2% 15% and 5%, 10% and 20% respectively. Mean has increasing average distance values as the missingness increases. This is intuitive as the number of missing values increase, the mean of the remaining values are biased towards larger values. All imputation schemes exhibit this trend. Figure 2b mirrors results



(a) Dice **(b)** Manhattan

Fig. 2. Normalized standard deviation distances of the 7 imputation methods using Lorentzian and Squared Euclidean distance

in Figure 1b whereby the lowest values are produced by EM for all levels of missingness. Generally, it is seen that the greater the level of missingness, the larger the average standard deviation.

Table 7. Normalized distances of the 7 imputation methods using Motyka distance

Method	1%	2%	5%	10%	15%	20%
EM	**0.120**	0.850	**0.168**	0.495	0.901	**0.062**
kNN	0.659	0.623	0.425	0.507	0.514	0.522
Mean	**0.120**	**0.145**	0.186	**0.272**	**0.353**	0.407
PCA	0.178	0.276	0.532	0.860	0.428	0.469
RF	0.192	0.195	0.238	0.314	0.384	0.421
SVD	0.216	0.574	0.312	0.330	0.365	0.429
SVT	0.505	0.461	0.787	0.503	0.486	0.502

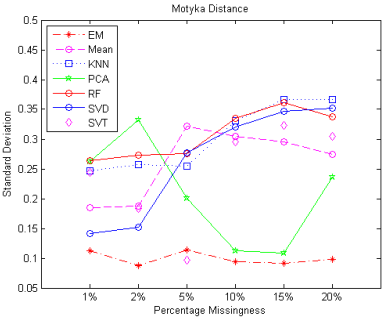


Fig. 3. Normalized standard deviation distances of the 7 imputation methods using Motyka distance

The results given by Table 7 slightly resemble those of Table 6 in that the lowest average distances are in favor of EM and Mean using the Motyka distance similarity. The minimum average distance (for EM and Mean) is 0.12 at the 1% level of missingness. The standard deviations of the Motyka similarity metric are presented in Figure 3. These values are similar to the Manhattan distance standard deviations (see Figure 2b) with the exception at the 5% level of missingness where the lowest standard deviation is produced by SVT. The average

standard deviations for all levels of missingness are less than 0.14 using EM, while the maximum for other imputation schemes exceed 0.3 (Mean).

6 Conclusion

In this paper, an investigation of the impact of imputation schemes on financial data was undertaken. A Monte Carlo simulation was done and randomly generated missingness (1%-20%) was induced having a ground truth dataset. Five distance metrics were used in order to measure the imputed datasets from the ground truth. The results show that using Squared Euclidean, Manhattan and Motyka similarity measures, EM generally achieves the closest distance to the benchmark data. Also the standard deviations using those metrics show that EM obtained the lowest score in general. This work is seen as an initial study. Future possible extensions include parameter tuning for the imputation schemes. Another important contribution would be to test whether imputed datasets outperform the benchmark with respect to classification accuracy.

Acknowledgments. The current work is being supported by the Department of Science and Technology (DST) and Council for Scientific and Industrial Research (CSIR).

References

1. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
2. Breiman, L.: randomforest: Breiman and cutlers random forests for classification and regression (2006)
3. Cai, J.F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization **20**(4), 1956–1982 (2010)
4. Crookston, N.L., Finley, A.O., et al.: yaimpute: An r package for knn imputation. Journal of Statistical Software **23**(10), 1–16 (2008)
5. Fogarty, D.J.: Multiple imputation as a missing data approach to reject inference on consumer credit scoring. Interstat (2006)
6. García-Laencina, P.J., Sancho-Gómez, J.L., Figueiras-Vidal, A.R.: Pattern classification with missing data: a review. Neural Computing and Applications **19**(2), 263–282 (2010)
7. Honaker, J., King, G., Blackwell, M., et al.: Amelia ii: A program for missing data. Journal of Statistical Software **45**(7), 1–47 (2011)
8. Jonsson, P., Wohlin, C.: An evaluation of k-nearest neighbour imputation using likert data. In: Proceedings of the 10th International Symposium on Software Metrics, 2004, pp. 108–118. IEEE (2004)
9. King, A.S., King, J.T.: The decision between debit and credit: finance charges, float, and fear. Financial Services Review **14**(1), 21–36 (2005)
10. Öğüt, H., Aktaş, R., Alp, A., Doğanay, M.M.: Prediction of financial information manipulation by using support vector machine and probabilistic neural network. Expert Systems with Applications **36**(3), 5419–5423 (2009)
11. Paleologo, G., Elisseeff, A., Antonini, G.: Subagging for credit scoring models. European Journal of Operational Research **201**(2), 490–499 (2010)

12. Regoeczi, W.C., Riedel, M.: The application of missing data estimation models to the problem of unknown victim/offender relationships in homicide cases. Journal of Quantitative Criminology **19**(2), 155–183 (2003)
13. Rezaee, Z.: Causes, consequences, and deterence of financial statement fraud. Critical Perspectives on Accounting **16**(3), 277–298 (2005)
14. Schafer, J.L., Graham, J.W.: Missing data: our view of the state of the art. Psychological Methods **7**(2), 147 (2002)
15. Sorjamaa, A., Lendasse, A., Severin, E.: Combination of soms for fast missing value imputation. In: Proceedings of MASHS, Models and Learnings in Human and Social Sciences (2010)
16. Zhou, W., Kapoor, G.: Detecting evolutionary financial statement fraud. Dezhoucision Support Systems **50**(3), 570–575 (2011)