Performance Analysis of Various Missing Value Imputation Methods on Heart Failure Dataset

Mohammad Al Khaldy(✉) and Chandrasekhar Kambhampati

Department of Computer Science, University of Hull, Hull, UK
m.a.al-khaldy@2014.hull.ac.uk,
C.Kambhampati@hull.ac.uk

Abstract. The missing data issue is a fundamental challenge in terms of analyses and classification of data. The classification performance of incomplete data could be affected and produce different accuracy results compared with complete data. In this work we compare six scalable imputation methods, implemented on a Heart Failure dataset. The comparison is done by the performance metrics of three different classification methods namely J48, REPTree, and Random Forest. The aim of the research is to find a classifier that achieves best performance results after imputing the missing data using different imputation methods. The results show that in general, the Random Forest classification achieves the best results in comparison to the decision tree J48 and REP Tree. Furthermore, the performance of classification improved when imputing the missing values by concept most common (CMC) and support vector machine (SVM).

Keywords: Heart failure · Decision tree · J48 · REPTree · Random forest · EM · Most common · CMC · KNN · K-mean · SVM

1 Introduction

Real life data often suffers from missing values. It is an important issue in the field of data mining since it can sometimes affect the classification accuracy, or the predictive modelling may be influenced by serious bias, besides, the bias occurs in the knowledge extraction [1–6]. Thus, the complete data is very important in terms of accurate use of the data and decision making processes [7]. There are many reasons why data may be missing, for example, the manual data entry procedure, data recorded and transferred errors, and incorrect measurement [1, 8]. Clinical data is a significant example of real data due to it playing an important role in discovering and implementing data mining and machine learning algorithms, clinical data can have serious effect on lifestyle, therefore it is important to manipulate and analyse this type of data.

Depending on the relation between the incomplete attributes and other completed attributes, there are three types of missing values [9]; (a) Missing Completely at Random (MCAR), (b) Missing at Random (MAR), and (c) Missing Not at Random (MNAR). MCAR occurs when the missing value is independent of other complete variables and does not depend on incomplete variables as well. That is, the attribute X

Y. Bi et al. (eds.), *Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016*,
Lecture Notes in Networks and Systems 16, DOI 10.1007/978-3-319-56991-8_31

being missing is not dependent on either complete data Y or X [10]. MAR occurs when the variable that is missing is independent from the values of that variable, but depends on other variables in the dataset, i.e. where the distribution is independent from missing itself but depends on other data [11]. The third mode, MNAR known as non-ignorable case, when it is not MCAR or MAR, i.e. probability of an instance being missing from an attribute depends on the value of that attribute [12].

The first phase of data mining and analysis is the pre-processing stage that includes removing noisy, feature selection, and manipulating missing values. There are two ways to deal with the missing values; deletion and imputation [3, 4, 11, 13, 14]. Deletion is used to discard all instances or variables with missing values, however, case deletion is not recommended due to the loss of data. The imputation is used to estimate or predict the missing values which use a model or mode/mean imputation. There are numerous model based imputation methods that can be categorised into implicit model based and explicit model based e.g. expectation maximization (EM) [14]. There are many methods of implicit model based such as K-nearest neighbor (KNN), weight KNN, Multi-Layer Perceptron (MLP), and RBFN. In addition, imputation can be categorised into single and multiple imputations (MI). Single imputation is a simple method such as mean, EM imputation. MI is more complex, generating a set of possible values for the missing value [15].

The objective of the research reported in this paper is to analyse and compare a set of imputation methods on various decision tree classifiers (C4.5, Random Forest, and REPTree). The dataset used is a clinical Heart Failure (HF) dataset. This analysis will help to find the most useful imputation method and the best decision tree classifier used to classify this kind of dataset.

The rest of the paper is organised as follows. Section 2 provides a brief literature review; in Sect. 3 a brief description of the clinical HF dataset is presented. Sections 4 and 5 describe the imputation methods and classification models used in the experiments. The details of the experiments are explained in Sect. 6. Finally, Sect. 7 draws the conclusion.

2 Related Work

Zhang et al. [8] presented a comparative study to find the most applicable imputation approach for the development of predictive models for the Heart Failure dataset. They examined several imputation methods and analyses their performance when applied to different classification algorithms. The results show that the missing values affect the classification process; also, handling missing values is significant in data mining processes.

Chauhan et al. [16] studied the intrusion detection classification to analyse network traffic data. They compared top-ten classification algorithms, where the comparison is based upon performance metrics. The experiments show that the decision tree classifiers are best for classifying the intrusions, especially the Random Forest algorithm.

Moore et al. [17] investigated how the underlying structures of the clinical data affects the performance of Bayesian classifiers. The results show that the imputation improves the performance of Naïve Bayes compared with SVM classifier.

Chau et al. [18] identified the heart disease of patients and compared the effectiveness and correction by using C4.5 with bagging, Naïve Bayes with bagging, and bagging algorithms, by computing the confusion matrix. Chau concluded that the best performance gained from the bagging was with Naïve Bayes.

Nakai et al. [19] categorised missingness with three cases from 5% to 50%, and studied the efficiency of four imputation methods (case method, mean method last observation carried forward, and multiple imputation method). Nakai concluded that multiple imputation is the most effective method.

3 Heart Failure Dataset

The dataset used is a real life heart failure dataset obtained from the Hull-LifeLab clinical database (University of Hull, UK). In this dataset, there are 61 variables with 1944 patient records. The missing values percentage in each feature ranges from 0% to 20%, while the missing values percentage of patient records is between 0% and 60%. The class is ('Alive' or 'Dead'), there are 485 'Dead' cases and 1459 'Alive' cases.

4 Missing Value Imputation Methods

Throughout the last two decades many imputation methods have been proposed. This research considers the most common model-based imputation methods.

4.1 K Nearest Neighbour Imputation (KNNI)

Nearest neighbour adopts the Euclidian distance to find the neighbours and then obtains a k cluster centre with neighbours together. The missing value instance is approximated by selecting the most similar instances [20]. K-NN is a lazy model, and its drawback is that this algorithm searches through all the dataset looking for the most similar instances, which is critical in the analysis of large datasets [3].

4.2 Expectation Maximization Imputation (EM)

EM is an iterative procedure involving two steps, Expectation (E-step) and Maximization (M-step), adopting maximum likelihood estimates for analysing complete data [21]. E-step uses the known attribute-value to estimate the parameters in the model for the data source, the previous iteration of the M-step and emission estimates of missing attributes values [22]. Then for each E-step the M-step maximizes the likelihood function to fill in the missing value [20].

4.3 K-Mean Imputation

This method finds the mean for a random *k* cluster. The dataset is divided into *k* groups based on similarity of objects and then fills the missing value by the mean of the group it belongs to. The similarity depends on the distance scale between the centre of the *k* cluster and the objects [23].

4.4 Most Common Imputation (MCI)

This method works simply by finding the most common attribute occurs is used to fill the missing value; this is when the data are symbolic. For numerical attributes, the missing values are filled by the average of all values in the attribute [24].

4.5 Concept Most Common Imputation (CMCI)

The method is similar to the Most Common but there is a restriction that uses cases belonging to the given class [25]. For the symbolic attributes, the missing value will be replaced by the most common attribute's value that occurs for the same concept, or replaced by the average of all values for the same concept [24].

4.6 Support Victor Machine (SVM) Imputation

Although the SVM algorithm is used for classification by recognising pattern and analyses data, it can be used to impute missing values. Adopting the chosen kernel, SVM defines the boundary between classes by selecting a set of support vectors [26]. The SVM is trained to use all examples in complete data training sets. The value of the variable imputed then becomes a target value and avoids the original classification. This method is iterative and while generating new training set, any other missing value attributes are ignored [27]. The training algorithm is slow and requires many complex computation processes [28].

5 Classification

The second part of this study involves the classification of a clinical dataset. The classification starts by training data for the target class. The decision tree is one of the main methods of learning a classification applied across a wide range of problems [29]. We chose the decision tree algorithms because they are the most commonly used techniques. The three decision trees selected here have different features. J48 is one of the most effective classification methods, REP Tree is a very fast algorithm, while Random Forest, although giving high accurate results, has a tendency to be very slow.

5.1 C4.5/ J48

Algorithm is an enhancement of the ID3 algorithm [30], the improvement of ID3 are features like speed, size, memory and a rule set output [31]. The algorithm steps are [32–34]:

(1) The tree represents a leaf for the instances existing in the same class, the leaf is returned by labelling with the same class.
(2) Calculate the information for each variable, selected by a test on the attributes, then calculate the achieved information that would be decided from a test on the variable.
(3) Iteratively apply the present selection principle to find the highest information gain; this variable is selected for branching.

5.2 REPTree

Reduced Error Pruning Tree (REPTree) is a fast decision tree learner that creates a decision tree adopting the information gain as the splitting criterion [35, 36]. The pruned tree reduced the pruning error and complexity [37].

5.3 Random Forest (RF)

RF consists of lots of decision trees based on a random selection of data and attributes. The X_n independent variables can be used for building a decision tree, the variables will be selected randomly into sets and these random decision trees create a forest [38]. The benefit of the large number of trees is that most of the trees can provide correct prediction of class. Another point is that all the trees do not make mistakes at the same place [39]. The final classifier gains accurate results since it is taken as a combination of more than one classifier [40].

6 Experiments and Results

The experiments were designed and implemented in a WEKA [41] environment for classification, and employed KEEL software [42, 43] to impute missing values for the heart failure dataset. Initially we conducted experiments to handle missing values in six different imputation methods (KNN, K-mean, CMC, SVM, Most Common, and EM). Then we conducted the imputed dataset on three different decision tree's classification (C4.5, Random Forest, and REPTree). After that, the performance of the classifiers is compared using the confusion matrix to find accuracy, sensitivity, specificity, in addition the execution time. The confusion matrix is a specific table, where each row represents the instances in a predicted class, with each column representing the instances in actual class (or vice-versa), as in Table 1.

Table 1. Confusion matrix

	Predicted NO	Predicted YES
Actual NO	True Negative (TN)	False Positive (FP)
Actual YES	False Negative (FN)	True Positive (TP)

The performance metrics can be measured by different equations such as, accuracy, sensitivity and specificity. Accuracy is simply measured by how the possibility that the algorithm can predicted negative and positive instances correctly [16] as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Sensitivity and specificity is the possibility that the algorithms can correctly predict positive and negative instances respectively, as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

The results in Table 2 illustrate the comparisons between the imputation methods by estimating the accuracy, sensitivity, and specificity for different decision tree classification algorithms. As can be seen, the imputation methods Concept Most Common (CMC) and Support Victor Machine (SVM) get the highest results with an accuracy of 85.71% and 84.97% respectively, when applying Random Forest for classification. For all imputation methods the results show that the classification methods are always ordered by: Random Forest, REPTree, and J48 descending. The table shows that the lowest accuracy and sensitivity are K-mean and Most Common Imputation, when applying the J48 algorithm for classification. The most critical note is that the classification without imputation gets accuracy and sensitivity greater than classification of the imputed data using Expectation Maximization and K-mean, see Figs. 1 and 2.

Figure 3 illustrates the time complexity for the classification methods used in this study. The REPTree algorithm has the highest speed in building the model while the Random Forest is the slowest algorithm in the decision tree methods. Because the REPTree algorithm sorts all numeric fields in the dataset once, and then uses the sorted lists to calculate the right splits in each tree node. The Random Forest is very slow because it creates too many trees and find many results then compare these results. From the figure also, we can see that the incomplete data take a long time to be classify compared with complete data that imputed by different imputation methods.

Table 2. Heart failure classification results. Accuracy, Sensitivity, and Specificity for the different classification algorithm used

Imputation method	Classification algorithm	Accuracy	Sensitivity	Specificity
EM	RF	78.05%	96.42%	22.88%
	J48	71.20%	81.56%	40.04%
	REPT	75.20%	90.55%	28.37%
K-mean	RF	78.76%	95.40%	28.66%
	J48	*70.88%*	*80.26%*	*42.68%*
	REPT	75.66%	91.36%	28.45%
Most Common	RF	78.28%	96.26%	24.25%
	J48	*68.97%*	*78.99%*	*37.44%*
	REPT	75.48%	91.24%	28.14%
CMC	RF	**85.71%**	**97.56%**	**50.11%**
	J48	81.03%	87.43%	61.78%
	REPT	82.57%	94.67%	46.22%
SVM	RF	**84.97%**	**97.10%**	**48.51%**
	J48	78.57%	91.24%	55.60%
	REPT	80.62%	92.76%	44.16%
KNN	RF	79.12%	95.68%	29.28%
	J48	71.14%	80.94%	41.65%
	REPT	74.38%	90.81%	24.95%
Without imputation	RF	78.18%	96.77%	22.26%
	J48	72.32%	84.37%	36.08%
	REPT	74.69%	91.08%	25.36%

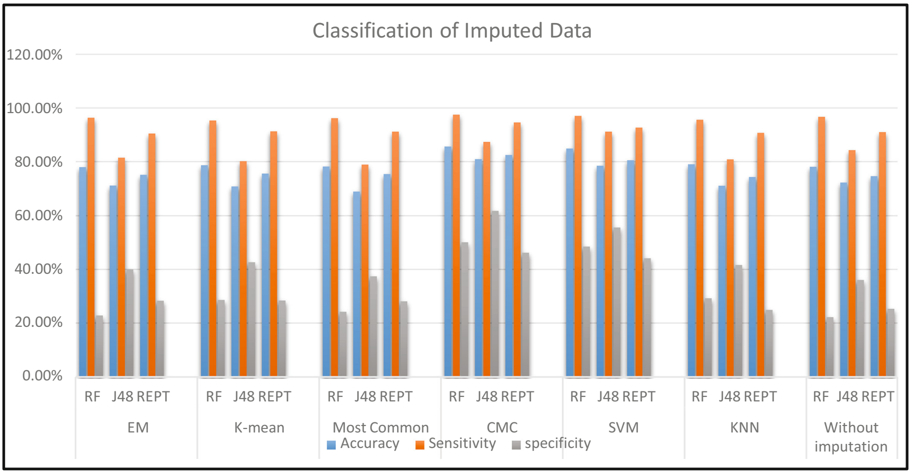

Fig. 1. The percentage of accuracy, sensitivity, and specificity using three classification algorithms on different imputation methods

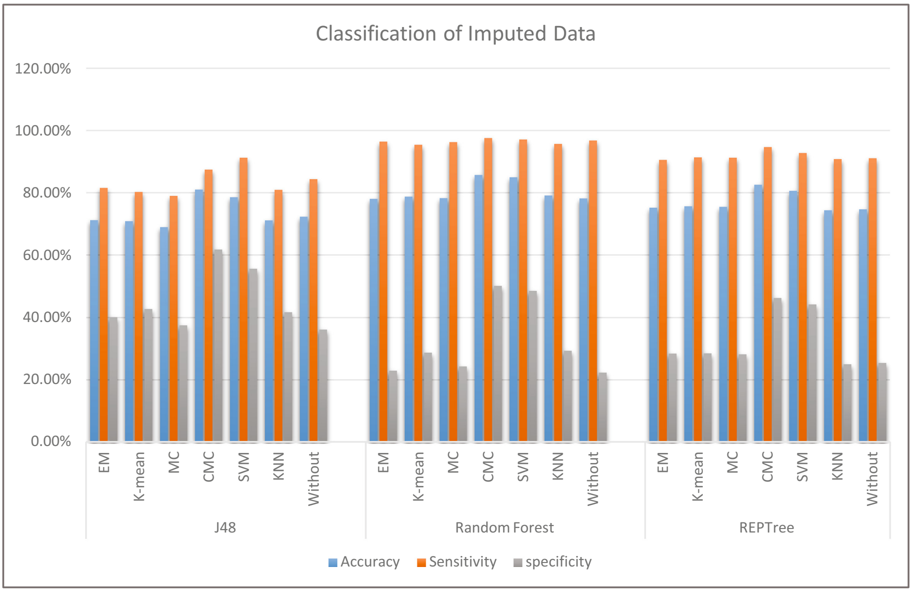

Fig. 2. The percentage of accuracy, sensitivity, and specificity using three classification algorithms on different imputation methods

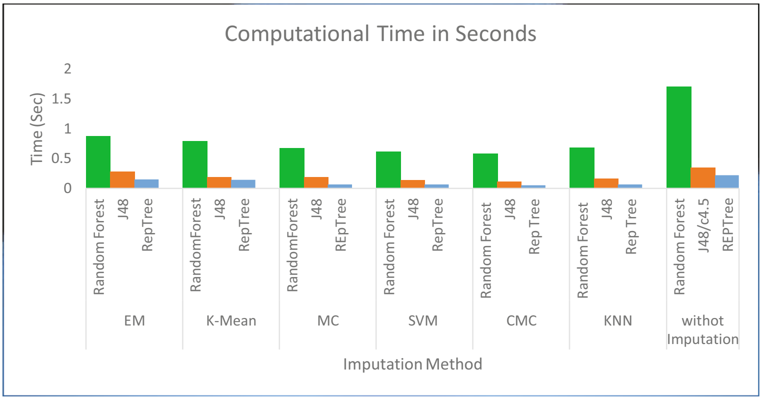

Fig. 3. The complexity time of different classification algorithms that used to classify the dataset imputed by different imputation methods

7 Conclusion

- In this paper, we investigated six different imputation methods and three different decision tree classification algorithms (Random Forest, REPTree, and J48). The imputation methods were employed on a Heart Failure dataset that contained 1944 instances and 61 attributes, where the data was suffering from missing values. After imputing the data it has been classified to estimate the classification performance in terms of accuracy, sensitivity, specificity, and computational time. From the experimental results, the following conclusions can be drawn:
- It is referred that concept most common (CMC) and support vector machine (SVM) outperform other imputation algorithms. On the other hand, expectation maximization and most common have minimum accuracy and sensitivity results.
- It is referred that Random Forest algorithm outperforms the other two decision trees, and J48 always gets minimum accuracy and sensitivity results compared with REPTree and RF.
- In term of processing time, it is referred that Random Forest is a very slow algorithm, while REPTree is very fast. On the other hand, all classification algorithms perform more slowly when classifying the dataset without imputation.
- Changing the value of K in KNN algorithm can change the accuracy and sensitivity results, and in our case the best output was when K = 6.
- Future work will be an investigation to implement and compare a deep multi-layer neural network to impute the missing values.

References

1. Liu, Z., Pan, Q., Dezert, J., Martin, A.: Adaptive imputation of missing values for incomplete pattern classification. Pattern Recogn. **52**, 85–95 (2015)
2. Razzaghi, T., Roderick, O., Safro, I., Marko, N.: Fast imbalanced classification of healthcare data with missing values. arXiv preprint arXiv:1503.06250 (2015)
3. Batista, G.E., Monard, M.C.: An analysis of four missing data treatment methods for supervised learning. Appl. Artif. Intell. **17**, 519–533 (2003)
4. Zhang, S., Qin, Z., Ling, C.X., Sheng, S.: "Missing is useful": missing values in cost-sensitive decision trees. IEEE Trans. Knowl. Data Eng. **17**, 1689–1693 (2005)
5. Marivate, V.N., Nelwamodo, F.V., Marwala, T.: Autoencoder, principal component analysis and support vector regression for data imputation. arXiv preprint arXiv:0709.2506 (2007)
6. Umathe, V.H., Chaudhary, G.: Imputation methods for incomplete data. In: 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1–4 (2015)
7. Carmona, C.J., Luengo, J., Gonzalez, P., del Jesus, M.J.: A preliminary study on missing data imputation in evolutionary fuzzy systems of subgroup discovery. In: 2012 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–7 (2012)
8. Zhang, Y., Kambhampati, C., Davis, D.N., Goode, K., Cleland, J.G.: A comparative study of missing value imputation with multiclass classification for clinical heart failure data. In: 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp. 2840–2844 (2012)

9. Little, R.J., Rubin, D.B.: The analysis of social science data with missing values. Sociol. Methods Res. **18**, 292–326 (1989)
10. Nelwamondo, F.V., Mohamed, S., Marwala, T.: Missing data: a comparison of neural network and expectation maximisation techniques. arXiv preprint arXiv:0704.3474 (2007)
11. Farhangfar, A., Kurgan, L., Pedrycz, W.: A novel framework for imputation of missing values in databases. IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum. **37**, 692–709 (2007)
12. Belanche, L.A., Kobayashi, V., Aluja, T.: Handling missing values in kernel methods with application to microbiology data. Neurocomputing **141**, 110–116 (2014)
13. Jordanov, I., Petrov, N.: Sets with incomplete and missing data—NN radar signal classification. In: 2014 International Joint Conference on Neural Networks (IJCNN), pp. 218–224 (2014)
14. Gheyas, I.A., Smith, L.S.: A neural network-based framework for the reconstruction of incomplete data sets. Neurocomputing **73**, 3039–3065 (2010)
15. Min, P.: Based on kernel function and non-parametric multiple imputation algorithm to solve the problem of missing data. In: 2011 International Conference on Management Science and Industrial Engineering (MSIE), pp. 905–909 (2011)
16. Chauhan, H., Kumar, V., Pundir, S., Pilli, E.S.: A comparative study of classification techniques for intrusion detection. In: 2013 International Symposium on Computational and Business Intelligence (ISCBI), pp. 40–43 (2013)
17. Moore, L., Kambhampati, C., Cleland, J.G.F.: Classification of a real live heart failure clinical dataset- Is TAN Bayes better than other Bayes? In: 2014 IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 882–887 (2014)
18. My Chau, T., Dongil, S., Dongkyoo, S.: A comparative study of medical data classification methods based on decision tree and bagging algorithms. In: 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, DASC 2009, pp. 183–187 (2009)
19. Nakai, M., Chen, D.-G., Nishimura, K., Miyamoto, Y.: Comparative study of four methods in missing value imputations under missing completely at random mechanism. Open J. Stat. **4**, 27–37 (2014)
20. Kumdee, O., Ritthipravat, P., Bhongmakapat, T., Cheewaruangroj, W.: Dealing with missing values for effective prediction of NPC recurrence. In: 2008 SICE Annual Conference, pp. 1290–1294 (2008)
21. Dodge, Y., Zoppe, A.: Adjusting the EM algorithm for design of experiments with missing data. In: 2004 26th International Conference on Information Technology Interfaces, vol. 1, pp. 9–12 (2004)
22. Karmaker, A., Kwek, S.: Incorporating an EM-approach for handling missing attribute-values in decision tree induction. In: 2005 Fifth International Conference on Hybrid Intelligent Systems, HIS 2005, p. 6 (2005)
23. Li, D., Deogun, J., Spaulding, W., Shuart, B.: Towards missing data imputation: a study of fuzzy k-means clustering method. In: Rough Sets and Current Trends in Computing, pp. 573–579 (2004)
24. Grzymala-Busse, J.W., Goodwin, L.K., Grzymala-Busse, W.J., Zheng, X.: Handling missing attribute values in preterm birth data sets. In: Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, pp. 342–351. Springer (2005)
25. Kaiser, J.: Dealing with missing values in data. J. Syst. Integrat. **5**, 42–51 (2014)
26. Sivapriya, T., Kamal, A.N.B., Thavavel, V.: Imputation and classification of missing data using least square support vector machines–a new approach in dementia diagnosis. Int. J. Adv. Res. Artif. Intell. **1**, 29–33 (2012)
27. Rogers, S.D.: Support vector machines for classification and imputation (2012)

28. Liu, Y., Liu, Y.: Incremental learning method of least squares support vector machine. In: 2010 International Conference on Intelligent Computation Technology and Automation (ICICTA), pp. 529–532 (2010)
29. Lomax, S., Vadera, S., Saraee, M.: A multi-armed bandit approach to cost-sensitive decision tree learning. In: 2012 IEEE 12th International Conference on Data Mining Workshops (ICDMW), pp. 162–168 (2012)
30. Agrawal, G.L., Gupta, H.: Optimization of C4.5 decision tree algorithm for data mining application. Int. J. Emerg. Technol. Adv. Eng. **3**, 341–345 (2013)
31. Sharma, P., Singh, D., Singh, A.: Classification algorithms on a large continuous random dataset using rapid miner tool. In: 2015 2nd International Conference on Electronics and Communication Systems (ICECS), pp. 704–709 (2015)
32. Kaur, G., Chhabra, A.: Improved J48 classification algorithm for the prediction of diabetes. Int. J. Comput. Appl. **98**, 13–17 (2014)
33. Almutairi, A., Parish, D.: Using classification techniques for creation of predictive intrusion detection model. In: 2014 9th International Conference on Internet Technology and Secured Transactions (ICITST), pp. 223–228 (2014)
34. Galathiya, A., Ganatra, A., Bhensdadia, C.: Classification with an improved Decision Tree Algorithm. Int. J. Comput. Appl. **46**, 1–6 (2012)
35. Mohamed, W.N.H.W., Salleh, M.N.M., Omar, A.H.: A comparative study of Reduced Error Pruning method in decision tree algorithms. In: 2012 IEEE International Conference on Control System, Computing and Engineering (ICCSCE), pp. 392–397 (2012)
36. Balasundaram, A., Bhuvaneswari, P.T.V.: Comparative study on decision tree based data mining algorithm to assess risk of epidemic. In: IET Chennai Fourth International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2013), pp. 390–396 (2013)
37. Junghun, P., Hsiao-Rong, T., Kuo, C.C.J.: GA-based internet traffic classification technique for qos provisioning. In: 2006 International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIH-MSP 2006, pp. 251–254 (2006)
38. Jian, X., Chen, P., Bin, L.: Random forest for relational classification with application to terrorist profiling. In: 2009 IEEE International Conference on Granular Computing, GRC 2009, pp. 630–633 (2009)
39. Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P.: Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. **43**, 1947–1958 (2003)
40. Cuzzocrea, A., Francis, S.L., Gaber, M.M.: An information-theoretic approach for setting the optimal number of decision trees in random forests. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1013–1019 (2013)
41. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. **11**, 10–18 (2009)
42. Alcalá-Fdez, A.F.J., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F.: KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. J. Multiple-Valued Logic Soft Comput. **17**(2–3), 255–287 (2011)
43. Alcalá-Fdez, J., Sánchez, L., García, S., Jesus, M.J., Ventura, S., Garrell, J.M., et al.: KEEL: a software tool to assess evolutionary algorithms to data mining problems. Soft Comput. **13** (3), 307–318 (2009)