Partially Synthesised Dataset to Improve Prediction Accuracy

(Case Study: Prediction of Heart Diseases)

Ahmed J. Aljaaf[1], Dhiya Al-Jumeily[1], Abir J. Hussain[1], Paul Fergus[1], Mohammed Al-Jumaily[2] and Hani Hamdan[3]

[1]Applied Computing Research Group, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, UK

[2]Dept. of Neurosurgery, Dr. Sulaiman Al Habib Hospital, Dubai Healthcare City, UAE

[3]CentraleSupélec, Département Signal & Statistiques, FRANCE

A.J.Kaky@2013.ljmu.ac.uk; {d.aljumeily, a.hussain, p.fergus}@ljmu.ac.uk; Hani.Hamdan@centralesupelec.fr

Abstract. The real world data sources, such as statistical agencies, library data-banks and research institutes are the major data sources for researchers. Using this type of data involves several advantages including, the improvement of credibility and validity of the experiment and more importantly, it is related to a real world problems and typically unbiased. However, this type of data is most likely unavailable or inaccessible for everyone due to the following reasons. First, privacy and confidentiality concerns, since the data must to be protected on legal and ethical basis. Second, collecting real world data is costly and time consuming. Third, the data may be unavailable, particularly in the newly arises research subjects. Therefore, many studies have attributed the use of fully and/or partially synthesised data instead of real world data due to simplicity of creation, requires a relatively small amount of time and sufficient quantity can be generated to fit the requirements. In this context, this study introduces the use of partially synthesised data to improve the prediction of heart diseases from risk factors. We are proposing the generation of partially synthetic data from agreed principles using rule-based method, in which an extra risk factor will be added to the real-world data. In the conducted experiment, more than 85% of the data was derived from observed values (i.e., real-world data), while the remaining data has been synthetically generated using a rule-based method and in accordance with the World Health Organisation criteria. The analysis revealed an improvement of the variance in the data using the first two principal components of partially synthesised data. A further evaluation has been conducted using five popular supervised machine-learning classifiers. In which, partially synthesised data considerably improves the prediction of heart diseases. Where the majority of classifiers have approximately doubled their predictive performance using an extra risk factor.

Keywords: Partially synthesised data; prediction, heart diseases, machine learning, rule-based method.

1 Introduction

There is growing interest from external researchers for access to data records collected by statistical agencies, organisations and research institutes. However, the privacy of individuals and confidentiality of data must be protected on legal and ethical grounds. Meanwhile, there is a demand to release a sufficient detail of data to maintain the reality and validity of statistical inference on the target population. To satisfy these desires, one method is to restrict data for approved analyses by authorised individuals. A second method is to release synthetic data rather than observed values, which typically conducted by a statistical disclosure control (SDC) technique [1]. The term of synthetic data has emerged since 1993 by Rubin [2]. The main aim was to protect the privacy and confidentiality of personal information through releasing synthetically produced data rather than actual data [2]. In general, a synthetic data can be created by a computer program using a random number generator or a formula that derived from real-world data [4]. There are two approaches of generating synthetic data, fully synthesis and partially synthesis data. Under the first approach, all data attributes are synthesised and no real data are released, while a subset of data attributes is synthesised under the partially synthesis approach [3, 5].

The real world data sources, such as statistical agencies, library databanks, research institutes and random generation procedures, are the major sources for researchers. Using this type of data involves a range of advantages. First, the data is relevant to real world problems, which enables reliable estimation of the usefulness of the results. Second, it improves the credibility and validity of the experiment. More importantly, this type of data is typically unbiased [4]. However, many studies showed that the use of synthetically generated data instead of real-world data is attributed to several factors including; a) the difficulty of using real-world data because of the privacy policies. (b) The available quantities of the real-world data may not be sufficient for the purposes of the experiment. (c) The collection of real-world data might be inapplicable, costly or time consuming. (d) The real-world data might be unavailable, particularly in the newly arises research subjects [3, 4]. Moreover, Synthetic data have a considerable advantages including; the simplicity of generation, requires relatively small amount of time in comparison with a real-world data collection, a sufficient quantity can be generated to fit the requirements with the diversity and relevance that can mimic the real-world data [4].

This study introduces a new method of creating synthetic data. We are proposing the generation of synthetic data from agreed principles using rule-based method. This new method has been proposed with the aim of improving prediction accuracy. In particular the prediction of heart diseases. We are targeting the improvement of heart diseases prediction through adding an extra risk factor. This risk factor will be synthetically generated in accordance with the World Health Organisation (WHO) criteria for classification of adults underweight, overweight and obesity according to BMI [20]. The experiment will be conducted using partially synthesised data, where more than 85% of the data have been extracted from real-world data, while less than 15% has been synthetically generated using rule-based method. The real-world part of data consists of six risk factors extracted from the Cleveland Clinic Foundation heart dis-

ease dataset, which available online at [16]. The synthesised part of data consists of one additional risk factor, which synthetically generated based on agreed principles. The Cleveland Clinic Foundation heart disease dataset was intensively used in the majority of studies that addressed the early prediction of heart diseases. These studies have used a full range of data attributes. In contrast, we are extracting only the risk factors, which represents the real-world part of data in this study. An adequate review of studies that targeted the prediction of heart diseases can be found in [6].

The researchers gave considerable attention to the prediction of heart diseases. Where the early prediction of heart disease has a significant influence on patient safety, as it can contributes to an effective and successful treatment before any severe degradation of cardiac output [6]. Heart diseases is a public health problem with high societal and economic burdens. It is considered the main cause of frequent hospitalisations in individuals 65 years of age or older, and slightly less than 5 million Americans suffer from heart diseases [8]. Heart diseases can occur because of many potential causes, some are illnesses in their own right, while others are secondarily to another underlying diseases [7, 8]. The commonest cause of heart failure is coronary disease by 62% compared to other risk factors such as hypertension, valvular disease, myocarditis, diabetes, alcohol excess, obesity and smoking [8, 9]. In general, heart diseases can be used to describe a condition in which the heart is unable to pump a sufficient amount of blood around the body [7].

In this paper, we aimed to a) review the latest studies that addressed the aspect of synthetic data generation, b) describe our proposed method of synthetic data generation using rule-based method and in accordance with agreed principles, c) inspect the feasibility of our method and adding an extra risk factor using principal component analysis, d) evaluate the utilisation of an extra risk factor to improve the prediction of heart diseases using five popular supervised machine-learning classifiers, and finally, e) highlight the results and study contributions.

2 Synthesised Data in Real-world Applications

Although the focus and the requirement are quite different in each field, the use of synthetic data has become an appealing alternative in many diverse scientific disciplines, including performance analysis, software testing, privacy protection and synthetic oversampling. Macia et al. [10] have proposed the use of fully synthesised data to investigate the performance of machine learning classifiers. As they have stated, the use of synthetically generated datasets can offer a controlled environment to analyse the performance of machine learning classifiers and therefore provide a better understanding of their behaviours. In the same context, Sojoudi and Doyle [11] have used a synthetic data generated by an electrical circuit model to investigate the performance of three methods, namely thresholding the correlation matrix, graphical lasso and Chow-Liu algorithm. These methods have been used as an alternative to identify the direct interactions between brain regions. They noticed that the first two

methods (i.e., thresholding the correlation and graphical lasso algorithm) are susceptible to errors.

In area of software evaluation, Whiting et al. [12] contributed in creating fully synthesised data to test visual analytics applications. Their main aim was to enable tool developers to determine the effectiveness of their software within an acceptable time frame. Similarly, Babaee and her colleague [13] have introduced the use of synthetic 2D X-ray images to validate medical image processing applications. Initially, a model of an organ is created using modelling software, then the model converted to computerised tomography image (CT) through assigning a proper Hounsfield unit to each voxel. As they have reported, this method may provide a ground truth data for researchers to validate their proposed medical image processing methods.

In another study, the researchers were targeting imbalanced learning problems. They have used a Kernel density estimation method to construct partially synthesised oversampling approach to address imbalanced class distribution in a particular data set. The experiment and evaluation was conducted using different medical related datasets with promising results [14]. Finally, Park and his partners have introduced a non-parametric systematic data for privacy protection of healthcare data. As they have claimed, their proposed method synthesises artificial records while maintaining the statistical features of the original records to the maximum extent possible. Using different data mining and statistical analysis methods, they have concluded that the synthetic dataset delivers results that largely similar to the original dataset [15].

3 Materials and Methods

3.1 Dataset

The experiment conducted using a partially synthesised data, which consists of two parts. First, real-world observations (i.e., risk factors), which represents 85.72% of the data. This part includes six risk factors extracted from the Cleveland Clinic Foundation heart disease dataset, which is available online at [16]. These risk factors are patient's age, gender, resting blood pressure, serum cholesterol, fasting blood sugar and maximum heart rate. This study selects these risk factors according to sophisticated researches in cardiovascular disease. As presented in the Framingham heart study, which predicting risk factors of cardiovascular disease for 30 years, patient's age, gender, blood pressure, serum cholesterol and blood sugar are considered standard risk factors [17]. Another long-term follow-up study on healthy individuals aged 25-74 years found that a high resting heart rate is an independent risk factor for coronary artery disease incidence or mortality among white and black individuals [18]. This part of data (i.e., real-world observations) consists of 297 consistent instances and without missing values. The output class includes four labels, which are no risk, low risk, moderate risk and high risk of developing heart diseases.

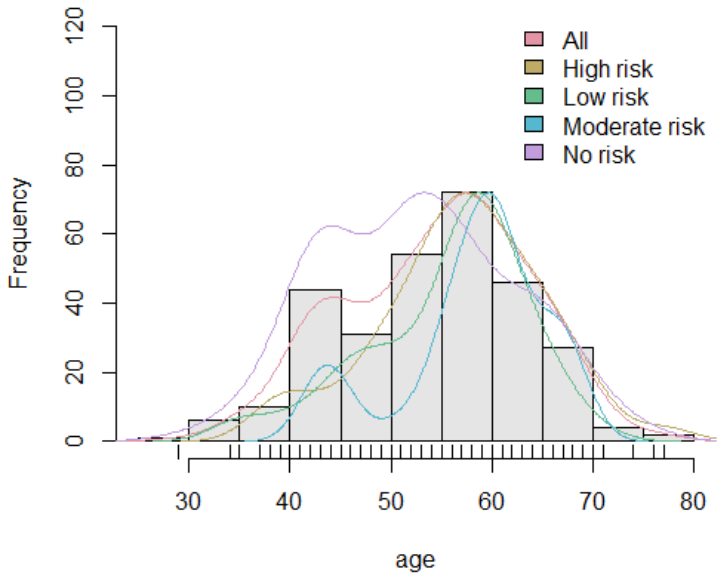

Figure 1. Distribution of age

Around two-thirds of the data are for male individuals, which is 201 instances. Mean age of individuals in the data set is 54 years. Figure 1 presents the age distribution, which clearly shows that the risk of developing heart disease starts approximately in the fourth decade of life. The risk is then about to double across every ten years to reach its peak in the sixth decade of life. Aging as shown by many studies poses the largest risk factor for cardiovascular diseases, where aging is associated with changes in cardiovascular tissues, which leads to the loss of arterial elasticity and increase arterial thickening and stiffness. These changes may subsequently contribute to hypertension, stroke, and arterial fibrillation [19].

The second part is an additional risk factor (i.e., body mass index BMI), which represents 14.28% of the data and has been synthetically generated in accordance with the World Health Organisation (WHO) criteria for classification of adults underweight, overweight and obesity according to BMI [20]. This study consider adding BMI as an additional risk factor because; a) it is neither been collected with the original data nor been involved with the same data for inference. b) The increase in BMI could dramatically increase the prospect of heart diseases. A study conducted in the USA showed that 30% reduction in the proportion of obese people would prevent approximately 44 thousand cases of heart diseases each year [7]. Moreover, being obese has been shown to double the risk of heart diseases [7, 8]. Finally, c) adding an additional risk factor would potentially improve the early prediction of heart diseases.

The second part (i.e., synthetic data) has been generated using rule-based method, in which, we are modelling the creation of WHO in the form of IF-THEN statements. These statements are implemented to generate the synthetic part of data. The classification of WHO identifies a principal cut-off points to categorise individuals according to their BMI, which is a manner of labelling someone as underweight, normal, over-

weight or obese. BMI is calculated by dividing an individual's weight in kilograms by the square of his/her height in metres [21]. As mentioned by WHO's global atlas on cardiovascular disease prevention and control, risks of coronary heart disease, raised blood pressure, type 2 diabetes and ischaemic stroke increase steadily with an increasing BMI [21]. Accordingly, the main aim of this paper is to involve BMI as an additional risk factor with the real world observation in order to improve the prediction accuracy of developing heart diseases. The following figure illustrates the distribution of the class labels in accordance with the cut-off points of the WHO criteria as a first step toward generating synthetic data from agreed principles.

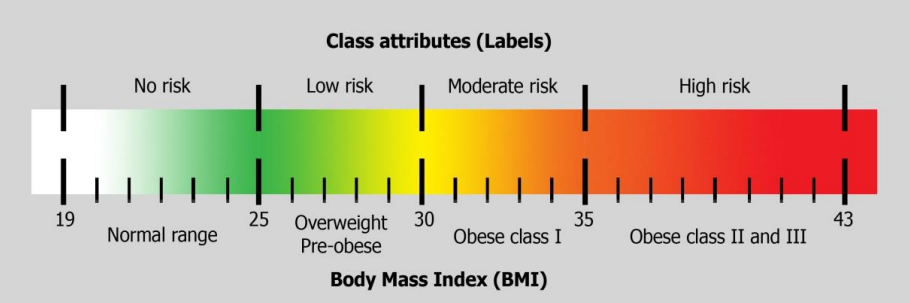

Figure 2. The distribution of class labels in accordance with BMI cut-off points

The first step was identifying the ranges of each class label. As shown in Figure 2, no risk class label assigned to normal range of BMI, which ranges from 19 and 25 according to WHO creation. Then low risk class label, which refers to overweight or pre-obese, represents BMI ranges between 25 and 30 according to WHO creation. Followed by moderate risk class label that corresponds to obese type one and finally high risk class label for obese type two and three. An overlapping area is maintained over the class labels, in which every class label is sharing a single unit with the following class label. For example, unit 25 of BMI is mutual between no risk and low risk class labels, and so on for the remaining class labels. This overlapping intended to simulate the real-world data disturbance and preserves statistical analysis from bias. The second step in the data generation process is to transform this explanation into conditional rules (i.e., IF-THEN statement). These rules then converted into a computer program that passes through the real-world dataset and generates BMI based on WHO creation and corresponding to a particular class label. The following algorithm demonstrates a second step of the data generation process, in which, we are using one SWITCH statement rather than a series of IF-THEN statement to express and translate the WHO criteria as a set of rules.

Input: a labelled set of real data
Output: a partially synthesised data

1. Begin
2. For each row ***r*** in the data file do
   2.1. Read class value ***clv***,
   2.2. Switch ***clv*** do
      2.2.1. Case ***clv*** : no risk
         Normal range of body mass index ***BMI***
         BMI = random number between (19, 25)
      2.2.2. Case ***clv*** : low risk
         A range from overweight to pre-obese of ***BMI***
         BMI = random number between (25, 30)
      2.2.3. Case ***clv*** : moderate risk
         A range of obese class one.
         BMI = random number between (30, 35)
      2.2.4. Case ***clv*** : high risk
         A range of obese class two and three
         BMI = random number between (35, 43)
   2.3. End switch
   2.4. Assign ***BMI*** value to the corresponding ***clv***
   2.5. End for
3. End

Algorithm 1. Partially synthesised data algorithm

Finally, let's consider A is patient's age, G is patient's gender, R is resting blood pressure, SC is serum cholesterol, FBS is fasting blood sugar and MHR is maximum heart rate. These risk factors are extracted for real world dataset. After adding the synthesised body mass index BMI, the partially synthesised dataset can be represented as a set $S= \{<A_1, G_1, R_1, SC_1, FBS_1, MHR_1, BMI_1>,...,< A_n, G_n, R_n, SC_n, FBS_n, MHR_n, BMI_n>\}$, where n is the size of the set S, which is the total number of instances. The Class attribute consists of four labels to classify patients with heart diseases into four risk levels and represented as $C = \{c_1,...,c_m\}$, in this study $m=4$, which are no risk, low risk, moderate risk, high risk.

3.2 Statistical Analysis

This section inspects the feasibility of involving the BMI as another risk factor for improving the prediction of heart diseases. A principal component analysis (PCA) is employed and the data is normalised. This experiment used a partially synthesised dataset, in which 85.72% of the data obtained from real-world data, while the remaining 14.28% have been synthetically generated using rule-based method. This combination of data belongs to different measurement scales including dichotomous (i.e., binary values) and continuous values. Two out of seven attributes are binary values,

which represent 28.57% of the data attributes. These attributes are patient's gender and fasting blood sugar. They were reported as 1 for male and 120 mg/dl or more of blood sugar, while zero for female and less than 120 mg/dl of blood sugar. The remaining five attributes are belonging to continuous values, but they have a different kind of distribution. For example, the mean and standard deviation of blood pressure attribute are 131.69 and 17.76, respectively, whereas they are 247.35 and 51.99 for serum cholesterol attribute. The mean of age attribute is 54.54 years.

These diverse ranges of means and standard deviations need to be transformed into a normal distribution before conducting statistical analysis, in which all the attributes will have the same mean or standard deviation. Therefore, the z-score normalisation method has been used to transform and unify the ranges of all attributes within the dataset. The z-score normalisation method re-scales the data to generate a new range dataset with zero mean and one standard deviation. The z-score normalisation method is mathematically represented as follows.

$$\bar{v}_i = \frac{v_i - \mu_x}{\sigma_x}$$

Where $\bar{v}_i$ is the normalised value, v_i is the original value, ith column of row attribute X, and μ_x and σ_x are the mean and standard deviation of row attribute X, respectively [22].

Although the principal component analysis (PCA) is commonly applied in data science for reducing the number of dimensions, particularly with high dimensional data, the main purposes of PCA is to identify patterns in data to highlight commonalities and variations. Therefore, PCA has been applied to describe the hidden structure of the data and investigate whether adding another risk factor maximising the amount of variance within the data or not. Where, maximising the amount of variance using the fewest number possible of principal components would be an ideal scenario of PCA, which reflect positively on the prediction accuracy. Figure 3 shows a score plot of the first principal component versus the second principal component of both real and partially synthesised data.

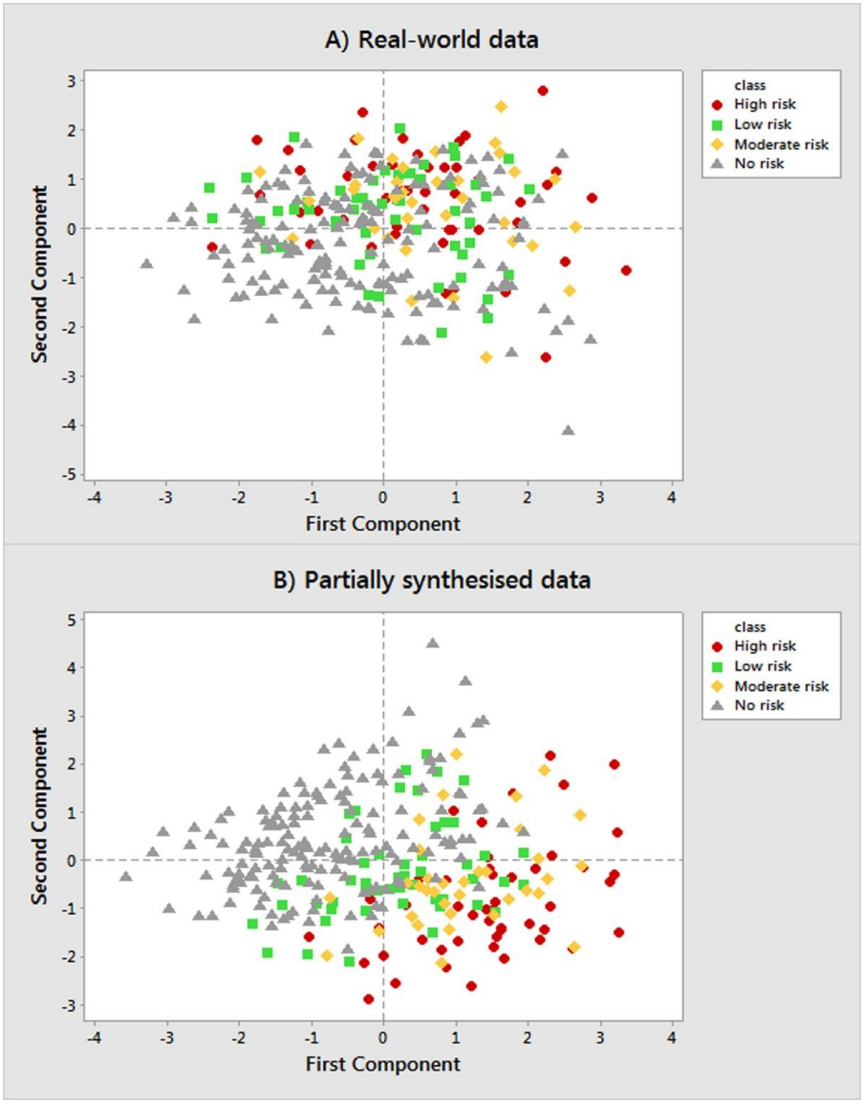

Figure 3. First principal component vs. second principal component

In Figure 3 A (i.e., real-world data), the score plot of the first two components shows a highly overlapping area among the class labels. Despite the cumulative proportion of eigenvalues of these two components should reveal an obvious groupings of data points, it appears difficult to aggregate the largest amount possible of data points that belong to a certain class label within a clear group. This is the leading cause at showing unsatisfactory results in the prediction of heart diseases. In contrast, with 14.28% synthesised data, the score plot of the first two components shows an improvement of the variance in the data. Figure 3 B (i.e., partially synthesised data), demonstrates a more separate distribution in the data points, which indicate that the

cumulative proportion of eigenvalues of these two components reveals an obvious groupings of data points. This will reflect positively on detecting clusters and improving predictions. Next section investigates this matter further through applying different machine learning methods.

3.3 Results

This section utilises five popular supervised machine-learning classifiers to assess both real and partially synthesised dataset, particularly evaluating the value of employing an extra risk factor to improve the prediction of heart diseases. The targeted classifiers are NaiveBayes (NB), Multilayer Perceptron Neural network (MLP), Support Vector Machine (SVM), Logistic Regression (LoR) and C4.5 Decision Tree (DT). This diversity of classifiers would clearly reveal whether an improvement in prediction of heart diseases from risk factors was accomplished using partially synthesised data or not. Two experiments have been conducted. In the first experiment, we have examined the sensitivity, specificity, mean absolute error and prediction accuracy of the targeted classifiers using the real-world data. In the second experiment, we re-examined these classifiers using partially synthesised data. In both experiments, we have used k-folds cross validation methods. In which, the data are partitioned into k equal subsets with almost the same proportions of different class labels. Of the k subsets, a single subset retained for testing, while the remaining k-1 subsets used for training. The cross validation is then repeated k times, until each subset applied exactly once for testing. Finally, the results averaged to estimate a model predictive performance. In this section, k=10.

Table 1. The evaluation of real-world data (i.e., risk factors)

Classifiers	Sensitivity	Specificity	Mean absolute error	Accuracy
NB	0.53	0.68	0.278	53.87
MLP	0.54	0.68	0.269	54.54
SVM	0.53	0.48	0.319	53.87
LoR	0.55	0.66	0.272	55.89
DT	0.51	0.70	0.271	51.85

Table 1 shows the overall predictive performance of the machine learning classifiers using real-world data. Results indicate a considerably low overall predictive performance, where the prediction accuracy ranges between 51% and 56%. Almost all the predictive models show convergent ranges of sensitivity, specificity and mean absolute error. This confirms the analysis results through the score plot of the first two components in Figure 3-A. Although unacceptable predictive performance (i.e., below average), Logistic Regression has achieved highest sensitivity and accuracy, followed by Multilayer Perceptron with 54% of sensitivity and accuracy. NaiveBayes has overcome Support Vector Machine with 2% of specificity; however, they have presented an identical sensitivity and accuracy. Despite Decision Tree comes at the end of the list with 51% of sensitivity and accuracy, it has recorded the best specificity. The

majority of predictive models showed approximately 0.27 of mean absolute error, except Support Vector Machine model, which recorded a slightly higher error rate.

Table 2. The evaluation of partially synthesised data (i.e., adding extra risk factor)

Classifiers	Sensitivity	Specificity	Mean absolute error	Accuracy
NB	0.91	0.95	0.064	91.58
MLP	0.90	0.97	0.052	90.90
SVM	0.87	0.92	0.260	87.20
LoR	0.93	0.98	0.037	93.93
DT	0.91	0.96	0.048	91.58

Table 2 introduces the overall predictive performance using partially synthesised data. The majority of classifiers have achieved impressive overall results with more than 90% of sensitivity, specificity and prediction accuracy. These results clearly demonstrate that the use of partially synthesised data (i.e., adding an extra risk factor) has had a significant impact on prediction accuracy. Logistic regression was also the leading model with 93% of sensitivity, accuracy and 98% of specificity. Although Multilayer Perceptron achieved the second highest specificity, its sensitivity and accuracy were slightly lower than NaiveBayes and Decision Tree that recorded similar sensitivity and accuracy. Support Vector Machine registered the lowest overall predictive performance. In contrast to the first experiment, the mean absolute error considerably dropped for the majority of models, whereas the overall predictive results substantially increased for all the predictive models.

3.4 Discussion

The conducted experiment highlighted the involvement of an additional risk factor, which synthetically generated using rule-based method and according to the standards of WHO, with a set of risk factors that extracted from real-world data to predict heart disease. Despite the lack of an agreed principle to indicate the accepted ratio of synthesised data in a particular data set, it seems that synthesising less than 15% of the data will not have a serious impact on the quality of statistical inference. In contrast, PCA has been conducted to investigate the hidden structure of the data, which reveals an improvement of the variance in the data using the first two principal components of partially synthesised data. This has been confirmed using various machine learning methods, where partially synthesised data significantly improves the prediction accuracy of heart diseases. The majority of classifiers have approximately doubled their predictive ability using the BMI as an extra risk factor.

This study holds two contributions. The main contribution shows the idea of generating synthetic data from agreed principles. A rule-based method has been used for this purpose. This strategy can be generalised into many other research areas. In particular the researches that aim to use fully and/or partially synthesised data in certain scientific discipline. For example, improving predictions, software testing and evalua-

tion, security aspects and so on. A reliable implementation of synthetic data generation using rule-based method requires a predefined criteria. Where these criteria need to be agreed worldwide in order to generate a valid set of data with a minimum possibility of bias. An expert knowledge is also under an obligation to express and translate these criteria into a set of rules. Where this study used the criteria of WHO with the aim of using an extra risk factor to improve the prediction of heart diseases. The second contribution is the participation of an additional risk factor to improve the prediction accuracy. This has been considered with several specialised studies [7, 8]. In contrast to the first contribution, the method of using extra risk factor to improve the prediction accuracy cannot be generalised as a new way to improve the prediction of certain diseases. It is entirely restricted to this study.

4 Conclusion

This paper presents the idea of generating synthetic data from agreed principles. The main aim was the improvement of the prediction of heart diseases from risk factors. Partially synthesised data have been used, in which more than 85% of the data extracted from real-world data, while the remaining was synthetically generated using rule-based method and in accordance with the criteria of World Health Organisation. A statistical analysis has shown an improvement in the variance of data after adding an extra risk factor. A further investigation has been conducted utilising five well-known supervised machine learning methods. The classifiers have approximately doubled their predictive performance using an extra risk factor, which confirms the statistical analysis result.

References

1. B. Loong, “Topics and Applications in Synthetic Data,” Doctoral dissertation, Harvard University, 2012.
2. D.B. Rnbin, “Discussion Statistical Disclosure Limitation,” Journal of Official Statistics, 1993; 9(3), pp. 461-468.
3. D.R. Jeske, B. Samadi, P.J. Lin, L. Ye, S. Cox, R. Xiao, T. Younglove, M. Ly, D. Holt, and R. Rich, “Generation of synthetic data sets for evaluating the accuracy of knowledge discovery systems,” in the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005, pp. 756-762.
4. N.G. Hall, and M.E. Posner, “The Generation of Experimental Data for Computational Testing in Optimization,” Chapter four in book entitled “Experimental Methods for the Analysis of Optimization Algorithms,” Springer, 2010.
5. J.W. Sakshaug, “Synthetic Data for Small Area Estimation,” Doctoral dissertation, The University of Michigan, 2011.
6. A.J. Aljaaf, D. Al-Jumeily, A. J. Hussain, T. Dawson, P. Fergus and M. Al-Jumaily, "Predicting the likelihood of heart failure with a multi level risk assessment using decision tree," in the Third International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), IEEE, Beirut, 2015, pp. 101-106.

7. The European Society of Cardiology, “Heart failure: Preventing disease and death worldwide,” Available at: http://www.escardio.org/communities/ HFA/Documents/whfa-whitepaper.pdf, Accessed in: 2 Feb. 2016.
8. VL. Roger, “The heart failure epidemic,” International Journal of Environmental Research and Public Health, 7(4), 2010, pp. 1807-1830.
9. Scottish Intercollegiate Guidelines Network (SIGN), “Management of chronic heart failure: A national clinical guideline,” Available at: http://sign.ac.uk/pdf/sign97.pdf, Accessed in: 5 Feb. 2016.
10. N. Macia, E. Bernado-Mansilla, and A. Orriols-Puig, "Preliminary approach on synthetic data sets generation based on class separability measure," in 19th International Conference on Pattern Recognition, (ICPR), IEEE, 2008, pp. 1-4.
11. S. Sojoudi and J. Doyle, "Study of the brain functional network using synthetic data," in 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2014, pp. 350-357.
12. M. A. Whiting, J. Haack, and C. Varley, "Creating realistic, scenario-based synthetic data for test and evaluation of information analytics software," Proceedings of the 2008 Workshop on beyond time and errors: novel evaluation methods for Information visualization, Florence, Italy, 2008.
13. M. Babaee and A. R. N. Nilchi, "Synthetic data generation for X-ray imaging," in 21th Iranian Conference on in Biomedical Engineering (ICBME), IEEE, 2014, pp. 190-194.
14. B. Tang and H. He, "KernelADASYN: Kernel based adaptive synthetic data generation for imbalanced learning," in the IEEE Congress on Evolutionary Computation (CEC), IEEE, 2015, pp. 664-671.
15. Y. Park, J. Ghosh, and M. Shankar, "Perturbed Gibbs Samplers for Generating Large-Scale Privacy-Safe Synthetic Health Data," in the IEEE International Conference on Healthcare Informatics (ICHI), IEEE, 2013, pp. 493-498.
16. The Cleveland Clinic Foundation, "Heart Disease Data Set," Available at: http://archive.ics.uci.edu/ml/datasets/Heart+Disease, Accessed in: 3 Feb. 2016.
17. M.J. Pencina, R.B. D'Agostino, M.G. Larson, J.M. Massaro, and R.S. Vasan, "Predicting the 30-Year Risk of Cardiovascular Disease: The Framingham Heart Study," Circulation, 2009; 119, pp. 3078-3084.
18. R.F. Gillum, D.M. Makuc, and J.J. Feldman, "Pulse rate, coronary heart disease, and death: the NHANES I Epidemiologic Follow-up Study," Am Heart J, 1991; 121, pp. 172–177.
19. B.J. North and D.A. Sinclair, "The Intersection between Aging and Cardiovascular Disease," Circulation Research, 2012; 110, pp. 1097-1108.
20. World Health Organisation, “The International Classification of adult underweight, overweight and obesity according to BMI,” Available at: http://apps.who.int/bmi/index.jsp?introPage=intro_3.html, Accessed in: 5 Feb. 2016.
21. The World Health Organization, “Global Atlas on cardiovascular disease prevention and control,” Available at: http://www.who.int, Accessed in: 3 Feb. 2016.
22. L. Al Shalabi, Z. Shaaban, "Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix," in the International Conference on Dependability of Computer Systems (DepCos-RELCOMEX 06), IEEE, 2006, pp. 207-214.