International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 11 (2018) pp. 9277-9282

Decision Support Predictive model for prognosis of diabetes using SMOTE and Decision tree

Shuja Mirza[1]*, Dr. Sonu Mittal[2], Dr. Majid Zaman[3].

[1, 2] Department of computer and system sciences. Jaipur National University Rajasthan.
[3] Directorate of IT&SS. University of Kashmir, Srinagar.

Abstract

Diabetes is one of the most appalling disease that mankind is facing currently. The disease occurs because of body's improper response to insulin: which is an important hormone in our body that converts sugar into energy needed for proper functioning of regular life. The diabetic disease has severe complications on our body as it increases the risk of developing kidney disease, heart disease, eye retinal disease, nerve damage and blood vessel damage. In this paper we developed a prediction model for diabetes prognosis using SMOTE and Decision tree classifier. Classification of imbalanced data especially in medical informatics is challenging. It was the main motivational factor for developing a classifier using SMOTE. We combined the two methods with an aim to improve the predictive accuracy of diabetic prognosis by removing class imbalance. The proposed system has two stages. In stage first the data imbalance is removed using SMOTE and in second stage the diabetes is diagnosed using Decision tree classifier. The obtained classification accuracy is 94.7013% and it was better as compared to other methods available in literature.

Keywords: Class imbalance, Diabetes, Decision tree, Medical diagnosis, Prognosis, SMOTE**.**

INTRODUCTION:

Diabetes, one of the metabolic disorders is the major threatening disease which has posed a great deal of threat to both developed as well as developing nations. The disease is characterized by high concentration of glucose in blood. The disease occurs by improper functioning of hormone known as "insulin" which acts as key to unlock cells allowing glucose to enter into it and fuel body with energy [1]. Diabetes is one of the leading causes of deaths worldwide especially in the developing nations it has wrecked havoc. As per statistics provided by WHO atleast 80% or more deaths occur in low and middle income countries as these countries lack adequate and high-end healthcare facilities. India too comes under category of developing countries and too has a huge number of diabetic patients to cater with and is considered as "diabetic capital of world" [2]. This major endocrine disease is found across populations and all age groups. In developed nations too the disease has caused lot of damage and is responsible for large number of deaths [3]. In 2014 fact sheets WHO has estimated about 422 million people suffering from diabetes with about 1.6 million deaths that were directly caused by diabetes and an estimate has been made that diabetes would be seventh leading cause of deaths by year 2030 [2]. As per diabetes atlas 2017 of IDF (International Diabetes Federation), by year 2045, an estimated 629 million people would be suffering with diabetes, with maximum number of people in age group 40-59. According to its statistics 1 in 2 persons *viz* approximately 212 million people are undiagnosed with diabetes. In year 2017 the diabetes caused the death of about 4 million people. The Asian belt has seen a steep rise in diabetes with about 138 million people affected by it (IDF) [4]. Diabetes comes with large number of complications which include kidney damage, blood vessels compression, and heart disease etc. what causes diabetes is still a mystery though studies have revealed two main factors genetic and environmental factors like obesity and unhealthy life style [5]. Diabetes has been categorized in two major types Type I (Juvenile) diabetes and Type II (adult onset diabetes). Adult onset diabetes or type II diabetes is the most common form of diabetes that accounts to about 90% of diabetic people suffering from it. Till date there has been no cure possible to treat diabetes although by changing the lifestyle and by necessary exercises the disease can be controlled. The challenge posed by the disease has compelled scientists for developing a prognostic DSS (decision support system) for aiding practitioners' in disease diagnosis.

In medical domain the popularity of data mining is increasing constantly as it helps exploring the unknown patterns and improves prediction models which help in medical decision making. In present times the health care all over the world is getting a great attention. The application of information technology has played a critical role in health care systems. In this current information age, the focal problem is how to deal with the huge amount of raw data that has been made available through distinct databases. In order to get the maximum amount of knowledge from these valuable databases, data mining techniques have been applied on it and these techniques have proven to be beneficial for early prediction of disease. Various data mining techniques have been used in medical domain for building predictive models for disease prediction, but with respect to medical science, classification had been an important decision making tool. The most popular techniques used are Decision trees, Support vector machines, Naïve Bayes.

As data mining methods acquire data from distinct sources, however the data in these datasets is often distorted. Majority of the real world datasets are imbalanced *viz*. majority of the instances are labeled to be belonging to one class called majority class while very few are labeled to be belonging to another class called minority class. This class imbalancing problem is prevalent in medical data. When such type of

imbalanced datasets is used in building of classifiers these methods tend to produce high accuracy for majority class and poor prediction for accuracy for minority classes [9, 10, 11]. A number of solutions have been proposed to handle imbalanced data both at data and algorithmic level, but SMOTE technique used in studies has shown better performance in literature.

In this study we have proposed a decision support predictive model based on SMOTE- dataset rebalancing technique and decision tree. We used SMOTE to reduce class imbalance. The obtained classification accuracy of this model has improved to 94.7% accuracy after rebalancing of dataset and shows potential with respect to other classification models in literature.

RELATED WORK:

The application of predictive classification in the field of medical diagnosis has received a great deal of attention, thanks to the strong research activities. In the recent past, the potential of predictive data mining to build clinically relevant decision support models from historical patient data have been highlighted by researchers. Distinct DSS (Decision Support Systems) built using different data mining algorithms have been introduced to assist medical experts, and each DSS is recognized by its accuracy. The prognosis of a particular disease through elevated level of accurateness has been the focal intend behind designing of DSS. Researchers have done a lot of research on Pima Indian data set to diagnose diabetes.

[12] Used various data mining algorithms that include C4.5, SVM, KNN for classification of diabetes, the highest accuracy of the result was that of C4.5 decision tree that had an accuracy of 86%. [13] In their research work applied Bayesian network for prediction of diabetes. They trained the dataset on this algorithm and obtained a predictive accuracy of 90.4%. [14] Applied genetic algorithm combined with fuzzy logic for prediction of diabetes. The model achieved an accuracy of 80.5%. [15] Used EM (Estimation Maximization], KNN, K-means, amalgam KNN and ANFIS on diabetic data set, however amalgam KNN and ANFIS achieved the highest accuracy rate of 80%. [16] In their research work used Multi Layer Perceptron, J48, and Naïve Bayes classification algorithms on Pima Indian dataset to predict diabetes, among three models Naïve Bayes achieved the accuracy of 76.30%. [17] developed a decision support system for diagnosis of diabetes; they used a combination of OLAP and data mining algorithm C4.5 and ID3 decision tree. The system achieved an accuracy of 74%. [18] Used decision tree and incremental learning for prediction of cardiac and diabetic symptoms, but model had an accuracy of just 68%. [19] Applied genetic algorithm on diabetic data for prediction of diabetes the model had an accuracy of 80%. [20] Proposed a model which they called "hybrid prediction model", the model was designed using simple K-means clustering algorithm and C4.5 decision tree classifier, the classification accuracy of the model was 92.38%. [21] Proposed an intelligent diagnosis system for diabetes prediction the system was based on Linear Discriminant Analysis (LDA) and Adaptive Network Based Fuzzy Inference System (ANFIS), the classification accuracy of this intelligent system was 84.61%. With the aim to diagnose diabetes [22] proposed a new cascade learning system. The system was based on GDA (General Discriminant Analysis) and LS-SVM (Least Square Support Vector Machine), with 10-fold cross validation the proposed system attained an accuracy of 82.05%. To detect diabetes [23] used PCA (Principal Component Analysis) and Adaptive Neuro-Fuzzy Inference System (ANFIS). The two methods were combined to improve prediction accuracy and achieved accuracy was 89.47%. [24] Designed a hybrid system to predict diabetes along with heart disease, to design such a system they used (ANN) Artificial Neural Network and (FNN) Fuzzy Neural Network, for diabetes prediction the model achieved a predictive accuracy of 84.24% using diabetic data. [25] Introduced an automated diagnosis system for prognosis of diabetes. The system was based on LDA (Linear Discriminant Analysis) and (MWSVM) Morlet Wavelet Support Vector Machine. The system had an accuracy of 89.47%.

CLASS IMBALANCE PROBLEM:

Class imbalance problem is presently one of the most sought-off topics in data mining research. The problem arises when number of instances of one class outnumbers another class. Most of the data mining algorithms that are used in medical diagnosis function well when supplied with evenly distributed class dataset, in majority of cases the supplied dataset is unevenly distributed viz, one class tend to have many instances than another thus leading to imbalance in dataset [26]. This predicament of imbalanced data is very much rife in medical dataset. This imbalance of classes in dataset causes many hindrances and difficulties' in the performance of machine learning and data mining techniques. As data mining methods attain knowledge from available diagnostic data and then this extracted knowledge is used for prognosis of a disease. But when this data is obtained from the source which is imbalanced viz, the class where majority of instances are labeled to as belonging to one particular class and few instances are labeled to be belonging to another, the situation can lead to poor predictive accuracy for minority class while as it will produce high accuracy for majority class. This state of affairs of imbalanced class distribution is the challenge posed for optimization of overall accurateness of classifier. In medical data mining when dataset is imbalanced the classifier tends to envisage high precision rate for majority class and tends to disregard minority one, but researchers preferences is for forecast of these minority classes with higher accuracy rate [27, 28]. To deal with this clause of class imbalance, researchers have turn up with two types of solutions, both at data as well as at algorithmic level. At data level different types of resampling techniques for instance, oversampling and undersampling can be employed, while as at algorithmic level resolution can be employed by applying design sophisticated classification approaches, the former method is preferential as it is straightforward to use [29, 30]. For enhancing forecast exactness and to beat the predisposition of model towards the greater part class due to data imbalancing we applied SMOTE (Synthetic Minority Oversampling Technique), it is an oversampling technique that work at data level as proposed by

Chawla et.al [31]. By applying SMOTE number of instances' in minority class can be increased by creation of new synthetic instances rather than by replication, thus it avoids overfitting problem in learning algorithms [31, 32, 33]. In SMOTE fresh instances are created "Synthetically" from minority class. SMOTE considers each one of instance as vector and generates synthetic samples alongside the line involving the minority sample and its nearest neighbor.

MATERIALS AND METHODS

Dataset used:

The dataset used in this investigate work is a clinical dataset. The dataset contains the record of 734 patients of just about all age groups. The dataset was taken from one of the leading diagnostic labs in Kashmir valley under the supervision of a medical doctor. During data compilation patients privacy and anonymity was appropriately taken care of, the composed dataset has 11 attributes as: age, plasma glucose fasting, post-Prandial glucose level taken 2-hours after meal, body mass index (BMI), systolic blood pressure, diastolic blood pressure, wais thickness, HbA1c value (3 months glucose concentration level), family diabetic history of patient and added classification attribute for indication of diabetic or non-diabetic. Table 1 has the description about dataset.

Table 1. Attributes used in our dataset.

Serial	Attribute name	Description	Values	Type
1	Age	Age of Patients in years	25-70 years	Numeric
2	Fasting	Fasting blood sugar level	75-115mg/dl	Numeric
3	Post_pran	Post Prandial blood sugar level	75-140mg/dl	Numeric
4	Waist	Waist measurement	30-40 inches	Numeric
5	BMI	Weight in kg's/height in m^2	20-40kg/m^2	Numeric
6	Systolic	Systolic blood pressure	90-170	Numeric
7	Diastolic	Diastolic blood pressure	60-100	Numeric
8	Hba1c	3 month plasma glucose concentration	3.5-10	Numeric
9	Gender	Gender of patient	M-male F-female	Nominal
10	History	Family history	1-yes 0-no	Numeric
11	Class	Diagnosis of disease	Yes No	Nominal

WEKA:

For implementation of our method we used WEKA (Waikato Environment for Knowledge Analysis) toolkit which was developed at university of Waikato New Zealand. This popular machine learning software is open source software and is freely available under GNU (General Public Licenses) [34]. The Weka tool contains number of standard machine learning methods which can be applied on datasets that are large enough to be analyzed manually for obtaining Knowledge. Weka understands data in ARFF (Attribute Relation File Format) format [35].

Decision tree:

Decision tree is amongst the most admired classification technique. Rules produced by the decision tree are 'anything but difficult to decipher and to understand' and hence it can help enormously in valuing the fundamental mechanism which separates' samples in different classes. Among a variety of decision trees based classifiers C4.5 algorithm is well-established and is extensively used. The algorithm uses information gain ratio criterion to make a decision about the most biased element at each progression of its decision tree induction process, with each round of determination, the information gain ration chooses the feature with maximum ration of its gain divide by entropy among features that had an average or better information gain. C4.5 stops building sub-tree when:

1. An obtained data subset contains samples of just single class.
2. No further feature is available (in that case leaf node is labeled as majority class).
3. When number of samples contained in obtained subset is less than a specified threshold (in that case leaf node gets labeled by the majority class) [36].

SMOTE:

It is an oversampling technique that was proposed by Chawla et.al [31], rather than operating in data space it operates in feature space. With the application of this approach the instances for minority class within the original dataset are increased, by creation of new synthetic instances, which are created with specification of two parameters 1 oversampling rate 2 number of nearest neighbors (k).

According to [26], new synthetic samples for continuous features are produced through following steps:

Step 1: by calculating of distance between a feature vector of minority class and one of its K-nearest neighbors.

Step 2: multiply the result obtained in step 1 by any random number in-between 0 and 1.

Step 3: addition of results in step 2 to the feature value of original feature vector.

We get a new feature vector as:

$$\gamma_{new} = \gamma_0 + (\gamma_{0i} - \gamma_0) \times \delta$$

γ_{new} represents the new synthetic sample.

γ_0 represents feature vector of each instance in minority class.

γ_{0i} represents i^{th} selected neighbor of γ_0.

ð represents random number between 0 and 1.

For nominal variables samples can be generated in following steps:

Step 1: takeout "majority vote" among features which are under deliberation and their K- nearest neighbor for nominal feature value, if clash or tie arises pick any arbitrary value.

Step 2: assigning the value obtained in step 1 to new synthetic minority class samples.

Take for example if we have a set of samples with features {P,Q.R,S.T} and its two nearest neighbor have set of features with samples as {P,Q,R,U,V} and {W,Q,R,U,V}, the new synthetic sample has set of features as {P,Q,R,S,V} [31].

EXPERIMENTATION

Two separate experiments were carried out on diabetic dataset by applying decision tree classifier. In first experiment we applied simple decision tree on dataset and obtained its result. In experiment second we applied decision tree combined with SMOTE for enhancing the classification accuracy of results. At the end the performance evaluation were carried out of both models and the obtained result was compared with some of the models available from literature.

Performance evaluation of proposed model

K-fold cross validation.

For evaluating the performance of our developed model, we used K-fold cross validation method for having a good measure of performance of the model [37]. The method divides the dataset into 'K' subsets and repeats the holdout method 'K' times, at each iteration one among 'K' subsets is used as test set while other K-1 subsets are grouped together forming a training set. At the end an average error of all 'K' iterations is calculated and that gives the test accuracy of our method. One of the main advantages if this method is that, how data is divided is of least importance.

Accuracy, Specificity, Sensitivity.

True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) all are the four achievable outcomes of a single prediction. True Positive (TP) and True Negative (TN) are correct classifications. These Classifications are given in form of Confusion Matrix. True Positive occurs when the sample is predicted positive and actually is positive. False Positive occurs when the sample is predicted as positive but actually is negative. True Negative occurs when sample is predicted as negative and actually is negative. False Negative occurs when sample is predicted as negative but actually is positive. For this study we used: 1. Accuracy. 2. Specificity. 3. Sensitivity. 4. Precision; equations for evaluation and analysis [37].

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

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$Precision = \frac{TP}{TP + FP}$$

RESULTS AND DISCUSSION

The mishmash of Decision tree and SMOTE on a diabetic set enhanced the prediction accuracy of the model as compared to model devoid of SMOTE. The decision tree classifier with SMOTE was classified, trained and tested on diabetic dataset. The obtained test classification accuracy of method was 94.2013% by using 10-fold cross validation which was an improvement over simple Decision tree with accuracy of 92.5068%. The accuracy of our model using combination of decision tree and SMOTE is higher than some of the models shown in table 5 taken from available literature. The obtained accuracy, sensitivity and specificity of the proposed model are 94.7013%, 94.4186% and 94.4013% respectively and are shown in table 2. The detailed accuracy of our model is given in table 4 while as table 3 depicts the overall error report. Our model had the ROC value of 0.953.

Table 2. Performance Measure

	Decision Tree	**Decision Tree with SMOTE**
Accuracy	92.5068%	94.7013%
Sensitivity	93.0232%	94.4186%
Specificity	91.7763%	94.4013%

Table 3. Error Report

Statistic	Decision Tree	Decision Tree with SMOTE
Kappa	0.846	0.8911
Mean absolute error	0.0784	0.0603
Root mean squared error	0.2555	0.218
Relative absolute error	16.1642%	12.4227%
Root relative absolute error	51.8785%	44.2653%

Table 4. Detailed Accuracy

Classifier	TP rate	FP rate	Precision	Recall	ROC
Decision Tree	0.925	0.077	0.925	0.925	0.955
Decision Tree with SMOTE	0.947	0.054	0.947	0.947	0.953

Table 5. Analysis of different data mining techniques for diabetes prediction

Technique	Sensitivity	Specificity	Accuracy	Reference
GDA-LS-SVM	82.05	83.33	79.16	[22]
LDA-ANFIS	83.33	85.18	84.61	[21]
PCA-ANFIS	85.71	92.01	89.47	[23]
C45.-K means	90.38	93.29	92.38	[20]
Our Model	**94.42**	**94.40**	**94.70**	**This study**

CONCLUSION

As there has been substantial enhancement in expert systems and machine learning tools, there effect has invaded more and more application domains with every passing day and medical field is no exemption. Decision making in medical field at times can be very troublesome. Classification systems used for making medical decision are provided with medical data which they examine in more comprehensive form but in shorter amount of time. In this research study we proposed a system that was based on Decision tree and SMOTE. We applied this system on diabetes diagnosis and one of the accurate learning mechanisms was evaluated. The study strongly suggests that reduction in class imbalance can result in improvement of prediction rate.

REFERENCES

[1] Mohamed, E. I., Linderm, R., Perriello, G., Di Daniele, N., Poppl, S. J., & De Lorenzo, A. (2002). Predicting type 2 diabetes using an electronic nose-base artificial neural network analysis. Diabetes Nutrition & Metabolism, 15(4),215–221.

[2] WHO. Facts and figures about diabetes. http://www.who.int/diabetes/en/

[3] Aljumah, A. A., Ahamad, M. G., & Siddiqui, M. K. (2013). Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University-Computer and Information Sciences, 25, 127–136.

[4] International Diabetes Federation. IDF Diabetes Atlas, 8th edn. Brussels, Belgium: International Diabetes Federation, 2017. http://www.diabetesatlas.org

[5] Polat, K., Gunes, S., & Aslan, A. (2008). A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. Expert Systems with Applications, 34(1), 214–221.

[6] Qian, Y., Liang, Y., Li, M., Feng, G., & Shi, X. (2014). A resampling ensemble algorithm for classification of imbalance problems. Neurocomputing, 143, 57–67.

[7] Ibarguren, I., Pe´rez, J. M., Muguerza, J., Gurrutxaga, I., & Ibarguren, O. A. I. (2015). Coverage based resampling: Building robust consolidated decision trees. Knowledge-Based Systems,. doi:10.1016/j.knosys.2014.12.023.

[8] Blaszczyn´ski, J., & Stefanowski, J. (2015). Neighbourhood sampling in bagging for imbalanced data. Neurocomputing, 150, 529–542.

[9] Sonu Kumari, Archana Singh "A data mining approach for the diagnosis of diabetes mellitus", IEEE conference on intelligent systems and control ,pp-373-375,2013.

[10] Rakseh Motka,Viral Parmar,"Diabetes Mellitus Forecast Using Different Data mining Techniques",IEEE International Conference on Computer and Communication Technology (ICCCT),2013.

[11] Veena Vijayan V., Aswathy Ravikumar, "Study of Data Mining algorithms for Prediction and Diagnosis

of Diabetes Mellitus", International Journal of Computer Application Vol 94,pp .12-16,June 2014.

[12] P. Radha, Dr. B. Srinivasan, " Predicting Diabetes by consequencing the various Data mining Classification Techniques", International Journal of Innovative Science, Engineering & Technology, vol. 1 Issue 6, August 2014, pp. 334-339.

[13] Mohtaram Mohammadi, Mitra Hosseini, Hamid Tabatabaee, "Using Bayesian Network for the prediction and Diagnosis of Diabetes" , MAGNT Research Report, vol.2(5), pp.892-902.

[14] Sudesh Rao, V. Arun Kumar, "Applying Data mining Technique to predict the diabetes of our future generations", ISRASE eXplore digital library, 2014.

[15] Veena vijayan, Aswathy Ravikumar, " Study of Data mining algorithms for prediction and diagnosis of Diabetes Mellitus", International Journal of Computer Applications (0975-8887) vol. 95-No.17, June 2014.

[16] Murat Koklu and Yauz Unal, " Analysis of a population of Diabetic patients Databases with Classifiers", International Journal of Medical,Health,Pharmaceutical and Biomedical Engineering", vol.7 No.8, 2013.

[17] Rupa Bagdi, Prof. Pramod Patil," Diagnosis of Diabetes Using OLAP and Data Mining Integration", International Journal of Computer Science & Communication Networks,Vol 2(3), pp. 314-322.

[18] Ashwinkumar.U.M and Dr. Anandakumar K.R, "Predicting Early Detection of cardiac and Diabetes symptoms using Data mining techniques", International conference on computer Design and Engineering, vol.49, 2012.

[19] S. Sapna, Dr. A. Tamilarasi and M. Pravin Kumar, "Implementation of Genetic Algorithm in predicting Diabetes", International Journal of computer science, vol.9 Issue 1, No.3, January 2012.

[20] Patil, Bankat M., Ramesh Chandra Joshi, and Durga Toshniwal. "Hybrid prediction model for type-2 diabetic patients." Expert systems with applications 37.12 (2010): 8102-8108.

[21] Dogantekin, Esin, et al. "An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: LDA-ANFIS." Digital Signal Processing 20.4 (2010): 1248-1255.

[22] Polat, Kemal, Salih Günes, and Ahmet Arslan. "A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine." Expert systems with applications 34.1 (2008): 482-487.

[23] Polat, Kemal, and Salih Günes. "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease." Digital Signal Processing 17.4 (2007): 702-710.

[24] Kahramanli, Humar, and Novruz Allahverdi. "Design of a hybrid system for the diabetes and heart diseases." Expert systems with applications 35.1-2 (2008): 82-89.

[25] Çalisir, Duygu, and Esin Dogantekin. "An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier." Expert Systems with Applications 38.7 (2011): 8311-8315.

[26] Chawla NV, Lazarevic A, Hall LO, Bowyer KW. SMOTEBoost:improving prediction of the minority class in boosting. In: KnowledgeDiscovery in Databases: PKDD 2003. Berlin, Heidelberg:Springer; 2003.

[27] He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–84.

[28] Sun Y, Wong AK, Kamel MS. Classification of imbalanced data: a review. Intern J Pattern Recognit Artif Intell. 2009;23(4):687–719.

[29] Y. Chen, Learning Classifiers from Imbalanced, Only Positive and Unlabeled DataSets, Department of Computer Science Iowa State University, 2009, pp. 1–5.

[30] M. Gao, X. Hong, S. Chen, C.J. Harris, A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems, Neurocomputing 74 (2011)3456–3466.

[31] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minor-ity over-sampling technique, Journal of Artificial Intelligence Research 16(2002) 321–357.

[32] Q. Gu, Z. Cai, L. Ziu, Classification of imbalanced data sets by using the hybrid re-sampling algorithm based on isomap, in: LNCS, Advances in Computationand Intelligence, 5821, 2009, pp. 287–296.

[33] T. Maciejewski, J. Stefanowski, Local neighbourhood extension of SMOTE formining imbalanced data, in: Proceeding of the IEEE symposium on computa-tional intelligence and data mining, 2011, pp. 104–111.

[34] Remco R. Bouckaert, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald and David Scuse "WEKA Manual for Version 3-6-8" THE UNIVERSITY OF WAIKATO, August 13, 2012.

[35] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

[36] Quinlan, J. R. (1993). C4.5 programs for machine learning. San Mateo, CA: Morgan Kaufmann Publishers

[37] Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. Artificial Intelligence in Medicine, 34, 113–127.