Clinical Epidemiology and Global Health 7 (2019) 160–164

Contents lists available at ScienceDirect

Clinical Epidemiology and Global Health

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

Data mining approach for accelerating the classification accuracy of cardiotocography

Sai Prasad Potharaju[a,*], M. Sreedevi[a], Vinay Kumar Ande[b], Ravi Kumar Tirandasu[b]

[a] Dept of CSE, K L University, Guntur, AP, India
[b] Dept of Computer Engineering, Sanjivani College of Engineering, Kopargaon, MH, India

ARTICLE INFO

Keywords:
Balanced
Imbalanced
Lazy learners
SMOTE
Rule based
Tree based

ABSTRACT

Objective: The objective of current study is to increase the classification accuracy of learning algorithms over cardiotocography data by applying preprocessing technique. Due to the diversity of sources, large amount of data is being generated and also has various problems including mislabeled data, missing values, noise, high dimensionality and imbalanced class labels.

Method: In this study, we suggested a technique to handle imbalanced data to increase the classification performance of various lazy learners, rule based induction models and tree based models. We used Symmetric Minority Over Sampling Technique (SMOTE) on real dataset to accelerate the performance of various classifiers. We identified that primary dataset is suffering with imbalanced problem, which means the most of the records belong to same class label. Prediction of imbalanced data is biased towards the class with majority instances. To overcome this problem, dataset has to be balanced.

Results: As a result of the suggested method the performance of classification algorithms are increased. The obtained result show that majority of classification techniques performed better over balanced dataset when compared with imbalanced dataset.

Conclusion: Classification performance over balanced dataset has recorded improved performance than imbalanced dataset after applying the SMOTE.

1. Introduction

Data mining (DM) is a process of discovering knowledge (interesting patterns) from huge datasets and is currently procuring extent deal of focus also became a prominent analysis tool.[1] In recent days, data mining techniques are applied in various fields such as stock market analysis, telecommunications, education institutes, human resource management, banking, supermarkets, health care management (HCM), traffic management and others. In data mining, classification and prediction are mostly applied for future planning and analysis of current trends. Data mining is a wider concept that contains different steps: Firstly data is pre-processed, where missing values are normalized, missing labels are rectified and noise will be minimized, then mining techniques (classification, association rule mining, clustering and others) are applied. Results from applied mining techniques are evaluated and interpreted.

For classification, in first phase construct the classifier, in which training dataset is used. Then, use the classifier for classification, in this phase testing dataset is used for calculating the performance of created classifier. In HCM, identifying the disease is critical and challenging job for the doctors. In recent days many countries are lighting on reasons which affect the patient's health. In recent days, due to the changes in lifestyle, health diseases are increasing. Especially, a disease related to heart has become more casual. There are no perfect analysis techniques existed to find unknown patterns in health care data. Data mining techniques can provide a solution in these situations. For this purpose, several data mining methods can be used. Thus, the aim of the current drive is to find out the class of fetal state of cardiotocograms from the collected patient dataset with the help of mining procedures.

Cardiotocograms (CTG) checks the fetal heart rate (FHR) and uterine shrinking, and are commonly applied as a diagnostic tool by obstetricians to know the fetal state.[2] Evaluation of CTG records need detailed visual interpretation subject. FHR checking is the process of monitoring the status of the baby during labor, and delivery by monitoring his or her heart rate with this special equipment called Cardiotocograms. During the process of monitoring the FHR, CTG will be kept on the stomach of the pregnant lady to monitor the child during birth. This is done to make sure the baby is delivered correctly.

* Corresponding author.
E-mail addresses: psaiprasadcse@gmail.com (S.P. Potharaju), msreedevi_27@kluniversity.in (M. Sreedevi), andevinay.gitam@gmail.com (V.K. Ande).

https://doi.org/10.1016/j.cegh.2018.03.004
Received 25 January 2018; Accepted 26 March 2018
Available online 27 March 2018

The road map of subsequent sections of the article is structured as follows: In second section, existing literature and related work regarding the paper is described. In third section, the suggested work of the current study is explained. Experimental evaluation of the suggested method is given in section four with the results. At the end, conclusion with future work is given.

2. Related work

There are number of research articles and methods have been contributed by many professionals on health care data. Research in health care ranging from: data collection, pre-processing, storage and analytics, by applying data mining strategies. In current days, research work on health care data has mainly focused on predicting variety of diseases, such as prediction of liver cancer, prediction of heart disease, prediction of diabetes, prediction of HCV, prediction of HIV and others. However, the focus of present study drives to evaluate the various methods of classification over the Cardiotocography Data Set.

A novel hierarchical fuzzy neural network approach is proposed to detect the breast cancer class as benign or malignant.[3] Clinical Decision System is proposed to get the medical expert using knowledgeable information.[4] Decision System proposed is built, based on weighted fuzzy rules.[4] Authors of,[5] presented a framework to detect various frauds in health care insurance using data mining methods. For this, authors considered historical insurance dataset.

To identify degree of liver fibrosis, the authors of,[6] proposed single and multi stage classification technique. Authors proved that, multistage model has given better result than existing techniques. To know the joint occurrences of diseases, association rule mining technique is applied on healthcare dataset.[7] Genetic K-Means Clustering is applied for effective health care knowledge discovery.[8] To predict the chances of heart disease based on various parameters like gender, blood pressure (BP), age, pulse rate, smoking habit, alcohol consumption and others are classified using Decision Tree (DT), Naive Bayes(NB), Artificial Neural Network (ANN).[9] For risk prediction of heart disease, various data mining techniques are applied in the research paper.[10]

Analysis of lung cancer is presented by applying different mining techniques by the authors of.[11] Authors of,[12] suggested a method to increase the prediction accuracy of kidney disease. Ensembling techniques proved the prediction of classification techniques better than the approach without ensembling.[13]

Support Vector Machine (SVM) performed well in classification of water quality status. For this SVM is compared with KNN algorithm.[14] Textual entries in health care records also classified by NB classification technique.[15] Computer based system is proposed to insert a needle in correct position during biopsy and other treatment.[16] NB, ANN, DT algorithms are used to predict the heart related disease.[17] Association rules can also used to predict the risk of diseases.Frequent itemsets are applied to predict the heart disease risk by considering the various symptoms.[18] Data mining techniques are not limited to only health care, this can be extended in many fields.Student performance and instructor performance are predicted using mining techniques.[19]

In literature, various classification algorithms are applied on different disease datasets. In this article, we used rule based, tree based and lazy learning algorithms for conducting our study. Table 1 describes the algorithms used in this study.

Table 1
Algorithms applied.

Category	Algorithms
Rule Based	Jrip, Ridor
Tree Based	J48,NB Tree
Lazy Learners	IBk, Kstar

Jrip is one of the popular rule based algorithm. It generates set of rules for a class. Ridor (Ripple Down Rule learner) is also a direct rule based classification method, which works in two phases. First, default rules are constructed, and then exceptions are produced for the rules generated by default with lowest error rate. Both these algorithms have the capabilities to handle nominal, binary, missing class values.

J48 is a classification technique used to build DT from a training dataset. It uses the concept of information gain for building DT. NBTree is also same as J48, it uses Naive Bayes classifiers for building DT. IBk implements the k-nearest neighbour algorithm. It is a lazy learner, which means model is generated for classification at the time of testing, Kstar is also a lazy learner and it is instance based classifier.

Due to the multiple heterogeneous platforms, large amount of data is being generated and also has following difficulties associated with it.

2.1. Mislabelled data

As data grows, the probability of having mislabelled records increases as well. When handling thousands of records, it is not easy to check whether all of the training data is correctly labelled or not, and training models on incorrect data will give less accuracy.

2.2. Missing values

Identical to mislabelled data, missing values also lead to inaccurate model generation when clustering algorithms are applied. This issue is generally minimized either by removing the instances completely or through imputation techniques.

2.3. Noise

Noisy data suffer from overfitting. Clustering techniques can assist to identify noisy data points.

2.4. High dimensionality

This problem happens when the features are more, or instances are very large. Principal Component Analysis (PCA) and Feature selection techniques can address this issue.

In this article also three filter based ranking feature selection methods namely: Chi squared attribute evaluator (Chi), Information gain (Ig), ReliefF attribute evaluator (Rel) applied on the data sets. These methods are based on the information theory. As per the information value associated with an attribute, rank will be assigned to each attribute. Depending on the requirement top 'N' features can be selected. Feature selection process is used to decrease the memory consumption and to increase the classification performance. Sometimes performance may be decreased depending on the dataset considered.[21]

2.5. Imbalance

In classification problems, imbalanced problem takes place in training data if more data points belongs to one class than in others. This problem can lead to weak learners.

This study addressed the issue of imbalance by using SMOTE.[20] It is an over sampling technique for balancing the imbalanced dataset. Using sampling technique the size of the dataset will be increased by adding synthetic instances. SMOTE uses K-nearest neighbour algorithm to increase the dataset. In this study SMOTE is the key concept which we used to improve the accuracy. For balancing the dataset, number of instances may be under sampled. This can be applied in the preprocessing stage of data mining.

Table 2
Dataset Description.[22]

S. No	Attribute Name	Type
1	LB	Real
2	AC	Real
3	FM	Real
4	UC	Real
5	DL	Real
6	DS	Real
7	DP	Real
8	ASTV	Real
9	MSTV	Real
10	ALTV	Real
11	MLTV	Real
12	Width	Real
13	Min	Real
14	Max	Real
15	Nmax	Real
16	NZeros	Real
17	Mode	Real
18	Mean	Real
19	Median	Real
20	Variance	Real
21	Tendency	Real
22	FHR pattern class code	Real
23	Fetal state class code	Categorical

Fetal State Class code has three values (1, 2, 3).
We replaced integer 1 with “One”, 2 with “Two”, 3 with “Three”.
Summary of Dataset is as follows.[21]
Instances: 2126.
#Attributes: 23.
#Classes: 3 (One, Two, Three).
Attribute Characteristics: Real.
Area: Life Sciences.
Associated Tasks: Classification.

3. Methodology

This section discusses the dataset description and suggested methodology for conducting the study.

For this study, we gathered dataset from UCI machine learning repository. The dataset belongs to the Cardiotocography and it has the measurements of FHR and uterine contraction (UC) features on CTG classified by expert obstetricians. Table 2 gives the description of dataset.

The dataset gathered is suffering with imbalanced problem. Initial dataset has 2126 instances and three classes namely One, Two, Three. Class label One has 295, class Two has 1655, class Three has 176 instances. It is clear that, class One and Three are not balanced with class Two.

Methodology of our system is given in Fig. 1.

To balance the complete dataset, 450% synthetic instances are created using SMOTE for class One, 750% synthetic instances are created for class Three. As a result of this process, total 4773 instances have been generated, out of these, Class label One has 1622, Two has 1655, Three has 1496 instances in new dataset (Balanced). Now the new dataset has almost balanced class labels. As SMOTE uses the K-Nearest Neighbour algorithm, we used K = 5 for sampling the dataset.

How an artificial instance will be created using SMOTE is explained with an example here.

Assume a datapoint (6, 4) and it’s closest datapoint is (4,3).

Let:

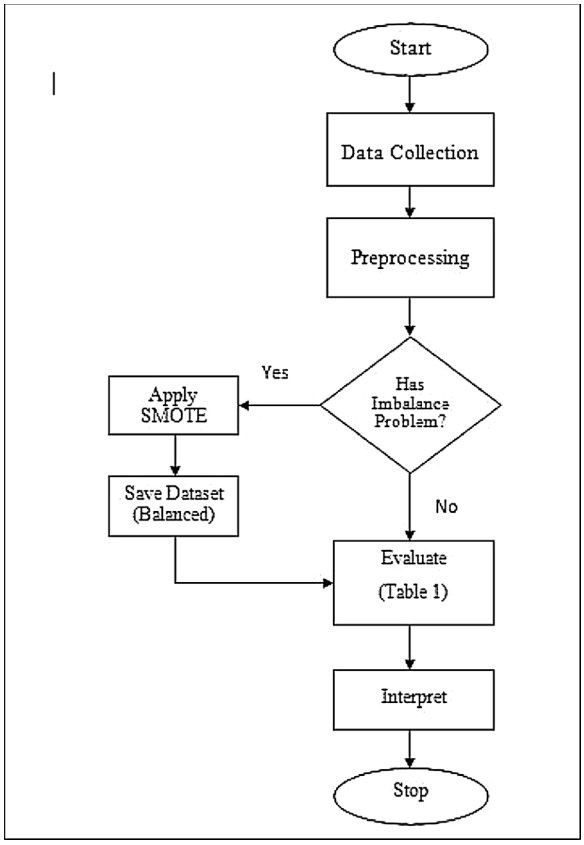

Fig. 1. Methodology.

f11 is the first attribute value of first data point.
f21 is the first attribute value of second data point.
f12 is the second(closest) attribute value of first data point.
f22 is the second(closest) attribute value of second data point.

Perform the $f_{new1}= f_{21}-f_{11}$
$f_{new2}= f_{22}-f_{12}$
As a result
$f_{new1}=(4-6)= -2;\ f_{new2}=(3-4)= -1;$
The new instance will be generated as per the below formula

$(f'_{new1,}\ f'_{new2})= (f_{11},f_{12})+Rand(0-1) * (f_{new1}\ ,\ f_{new2})$
$(f'_{new1,}\ f'_{new2})= (6,4)+Rand(0-1) * (-2,-1)$
$=(6,4) +(0.1)*(-2,-1)$
$=(6,4)+(-0.2,-0.1)$
$=(5.98,3.99)$ is the new instance to be created .

Note: Rand(0-1) produce the random number between 0 and 1.It can be minimum and maximum value of the respective attribute.

After applying the SMOTE over the imbalanced dataset, some of the feature selection techniques (Chi, Ig, Rel) are applied over the balanced and imbalanced datasets. Those results are given in next section also.

4. Experiment and results

For experiment the suggested methodology, data mining tool weka is used. For evaluation of balanced and imbalanced dataset, algorithms given in Table 1 are applied. System configuration for the experiment is: operating system:Ubuntu 14.04 LTS, processor: Intel® Core™ i5 CPU M 430 @ 2.27 GHz × 4, memory:6 GB.

Dataset is divided into 2:3 and 1:3 ratios for training and testing respectively. 10 fold cross validation is applied for calculating the accuracy of algorithms. Table 3 shows the accuracy and ROC of algorithms over the imbalanced and balanced datasets.

Table 3
Accuracy of classifiers over imbalanced and balanced datasets.

	Imbalanced		Balanced	
	Accuracy	ROC	Accuracy	ROC
Jrip	**98.73**	0.978	98.38	0.992
Ridor	98.30	0.97	98.36	0.988
J48	98.58	0.984	98.78	0.993
NBTree	97.31	0.981	97.98	0.995
IBK	96.89	0.963	**99.05**	0.993
KStar	94.63	0.987	97.86	0.998

Bold Indicates the highest accuracy.

Table 4
Average result of rule based, tree based and lazy learning algorithms.

Dataset Type	Rule based classifiers	Tree based Classifiers	Lazy Learners
Imbalanced	98.51	97.94	95.76
Balanced	98.37	**98.38**	**98.45**

Bold Indicates the highest accuracy.

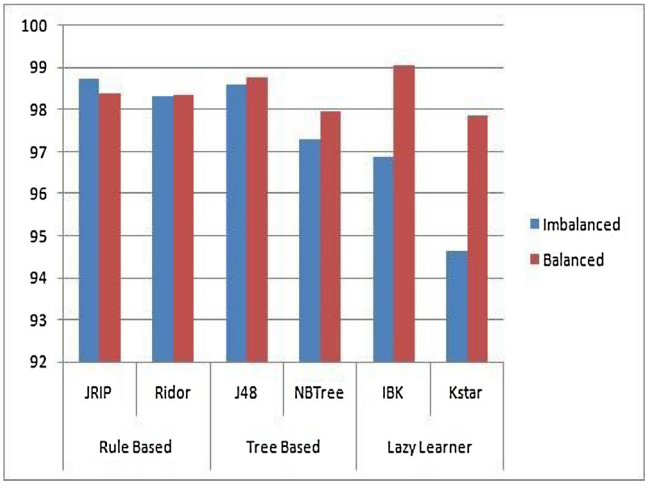

Fig. 2. Comparison of algorithms over Imbalanced and Balanced datasets.

From the above table, Jrip has recorded best performance over imbalanced data and IBK performed better than other classifiers over balanced dataset. Fig. 2 shows the comparison of algorithms over imbalanced and balanced datasets

From Fig. 2, it is clear that, rule based Ridor, tree based J48 and NBTree, and lazy learners IBK and Kstar performed well in case of balanced dataset. In other way rule based Jrip has performed well in the case of balanced data. Table 6 shows the average result of rule based, tree based and lazy learning algorithms.

From the above result, the average performance of tree based classifiers and lazy learners are better over balanced dataset. Fig. 3 shows the average comparison of rule based, tree based and lazy learning algorithms.

In Figs. 2 and 3 X-axis represents the type of classifier,Y-axis represents the percentage of accuracy.

After analyzing the various classifiers over balanced and imbalanced datasets, three feature selection methods(Chi, Ig, Rel) are applied. Actual dataset has 22 features in it (Refer Table 2). As these feature selection methods assigns the rank to each feature, approximately 30% features(top 6) are selected for classification from the imbalanced and balanced datasets. The list of top features derived by feature selection methods are given in Table 5. Classification performance with those features is given Table 6.

From Table 5, it can be observed that Chi, Ig, Rel feature selection

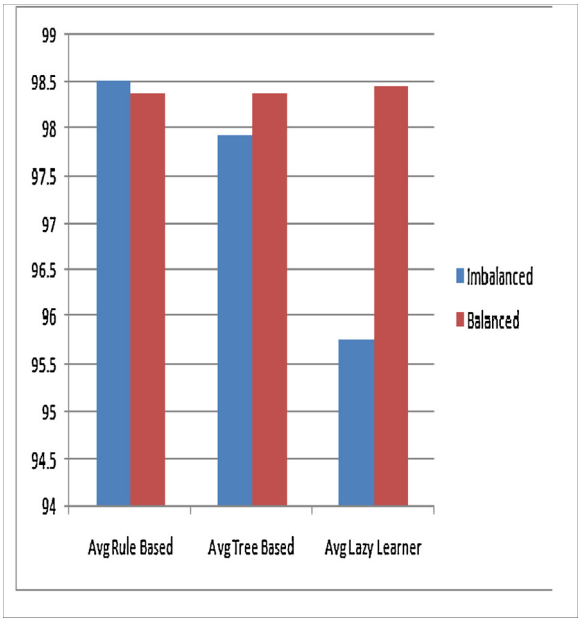

Fig. 3. Average comparison of rule based, tree based and lazy learning algorithms.

Table 5
Features derived by feature selection methods.

Type	Feature selection Method	Selected Feature id's
Imbalanced	ChiSquaredAttributeEval (Chi)	22,10,18,9,8,17
	Information Gain(Ig)	22,9,10,8,18,20
	ReliefFAttributeEval (Rel)	22,8,10,2,4,1
Balanced	ChiSquaredAttributeEval (Chi)	22,9,17,18,19,10
	Information Gain(Ig)	22,9,17,18,19,10
	ReliefFAttributeEval (Rel)	22,8,10,4,18,1

Table 6
Classification Performance after Feature Selection Methods.

Dataset Type	FS Method	Jrip	Ridor	J48	NBTree	IBK	KStar
Imbalanced	CHI	98.91	98.35	98.73	98.25	98.07	96.89
	Ig	98.77	98.54	98.77	98.4	97.78	97.36
	Rel	98.68	98.54	98.77	98.58	97.46	96.84
	ALL	98.73	98.3	98.58	97.31	96.89	94.63
Balanced	CHI	98.11	97.67	98.3	97.17	97.52	97.1
	Ig	98.11	97.67	98.3	97.17	97.52	97.1
	Rel	98.14	97.98	98.13	97.8	98.44	97.42
	ALL	98.38	98.36	98.78	97.98	99.05	97.86

Note: 'ALL' indicates the classification accuracy by considering the whole dataset.

methods derived the different features over imbalanced dataset. Chi and Ig derived same features over balanced dataset. Classification performance with those features is given in Table 6.

From the above Table 6 following outcomes are observed. Over imbalanced dataset, Jrip and IBK classifier produced maximum performance with Chi feature selection methods than considering all the features. Ridor and J48 displayed highest accuracy with Ig and Rel. NBTree recorded better performance with the Rel feature selection method. Kstar classifier is recorded better performance with Ig than considering all the features.

Performance of algorithm's accuracy is calculated using confusion matrix. Basic structure of confusion matrix is given in below Table 7.

Table 7
Confusion Matrix.

		Predicted	
Actual		Class Yes	Class No
	Class Yes	TP	FN
	Class No	FP	TN

Where: TP is True Positive.
TN is True Negative.
FP is False Positive.
FN is False Negative.
Accuracy: (TP + TN)/(Total number of instances).

5. Conclusion

In this article, we analyzed Cardiotocography dataset for classification of fetal state class using Jrip, Ridor, J48, NBStar, IBk, and Kstar. Initially dataset is imbalanced. So, by applying SMOTE, dataset has balanced. Then, above said techniques are applied on both the datasets. Experimental result shows that, classification performance on balanced dataset has recorded improved performance than imbalanced dataset. Three feature selection methods also applied to analyze the performance after selecting top 6 features. This technique can be applied in case of huge amount of data(Big data) with Map Reduce technique, as the traditional mechanism does not fit to handle Big Data.

References

1. Data mining applications for empowering knowledge societies. In: Rahman H, ed. Information Science Reference: Hershey, PA; 2009.
2. 2.Johnson B, Bennett A, Kwak M, Choi A. Automated evaluation of fetal cardiotocograms using neural network. *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on.* 2012:408–413.
3. Seyedeh S, Mohammad T, Mahdi A. Breast cancer detection by using hierarchical fuzzy neural system with EKF trainer. *Proceedings of the 17th Iranian Conference of Biomedical Engineering (ICBME2010).* 2010; 2010.
4. Anooj PK. Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules. *J King Saud Univ – Comput Inf Sci.* 2012;24(1):27–40. http://dx.doi.org/10.1016/j.jksuci.2011.09.002.
5. Kirlidog M, Asuk C. A fraud detection approach with data mining in health insurance. *Procedia – Social Behav Sci.* 2012;62:989–994. http://dx.doi.org/10.1016/j.sbspro.2012.09.168.
6. Hashem AM, Rasmy MEM, Wahba KM, Shaker OG. Single stage and multistage classification models for the prediction of liver fibrosis degree in patients with chronic hepatitis C infection. *Comput Methods Prog Biomed.* 2012;105(3):194–209. http://dx.doi.org/10.1016/j.cmpb.2011.10.005.
7. Rashid MA, Hoque MT, Sattar A. *Association rules mining based clinical observations. arXiv preprint arXiv:1401.2571.* 2014; 2014 Retrieved from https://arxiv.org/abs/1401.2571.
8. Alsayat A, El-Sayed H. Efficient genetic K-means clustering for health care knowledge discovery. *Software Engineering Research, Management and Applications (SERA), 2016 IEEE 14th International Conference on.* 2016:45–52.
9. Thomas J, Princy RT. Human heart disease prediction system using data mining techniques. *Circuit, Power and Computing Technologies (ICCPCT), 2016 International Conference on.* 2016:1–5.
10. Purusothaman G, Krishnakumari P. A survey of data mining techniques on risk prediction: heart disease. *Indian J Sci Technol.* 2015;8(June (12)):2015.
11. Venkatdass M, Mohammed AR, Mohammed A. Classification of lung cancer subtypes by data mining technique. *International Conference on Control, Instrumentation, Energy & Communication(CIEC), 2014.* 2014.
12. Prasad Potharaju S, Sreedevi M. An improved prediction of kidney disease using SMOTE. *Indian J Sci Technol.* 2016;9(31)http://dx.doi.org/10.17485/ijst/2016/v9i31/95634.
13. Potharaju SP, Sreedevi M. Ensembled rule based classification algorithms for predicting imbalanced kidney disease data. *J Eng Sci Technol Rev.* 2016;9(5):201–207.
14. Danades A, Pratama D, Anggraini D, Anggriani D. Comparison of accuracy level K-nearest neighbor algorithm and support vector machine algorithm in classification water quality status. *System Engineering and Technology (ICSET), 2016 6th International Conference on.* 2016:137–141.
15. Aldhoayan M, Zhou L. An accurate and customizable text classification algorithm: two applications in healthcare. *Computational Advances in Bio and Medical Sciences (ICCABS), 2016 IEEE 6th International Conference on.* 2016:1–4.
16. Yang Y, Ahmed A, Yue S, Xie X, Chen H, Wang Z. An algorithm for accurate needle orientation. *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the 5095–5098.* 2016.
17. Gandhi M, Singh SN. Predictions in heart disease using techniques of data mining. *Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015 International Conference on.* 2015:520–525.
18. Ilayaraja M, Meyyappan T. Efficient data mining method to predict the risk of heart diseases through frequent itemsets. *Procedia Comput Sci.* 2015;70:586–592. http://dx.doi.org/10.1016/j.procs.2015.10.040.
19. Shahiri AM, Husain W, Rashid NA. A review on predicting student's performance using data mining techniques. *Procedia Comput Sci.* 2015;72:414–422. http://dx.doi.org/10.1016/j.procs.2015.12.157.
20. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority oversampling technique. *J Artif Intell Res.* 2002;16:321–357.
21. Potharaju SP, Sreedevi M. A novel M-cluster of feature selection approach based on symmetrical uncertainty for increasing classification accuracy of medical datasets. *J Eng Sci Technol Rev.* 2017;10(6):154–162.
22. https://archive.ics.uci.edu/ml/datasets/Cardiotocography#.