Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Stacked Approach in Medical Data Mining for the Post Operative Patient's Recovery Area Prediction

[1]Chinky Gera, [2]Kirti Joshi
[1]Research Scholar, [2]Assistant Professor
[1, 2] Department of Computer Science and Engineering
RIMT Institute of Engineering and Technology, PTU, Punjab, India

Abstract— Patient's post operative data is allied with hypothermia which is a serious concern after surgery. To operate against hypothermia is crucial for patient's health. In this paper, various techniques of data mining are considered for the prediction of post operative patient's data. The research affirms that the testing has been done on the test dataset containing 13 instances with unknown classes on the training dataset in which stacked technique with three different classifiers has outperformed using resample filter of supervised approach to remove the biasness among classes. By using Weka 3.7.4, experimental result shows that stacked technique achieves 94.67% of accuracy.

Keywords— Data mining, Post operative patient, Stacked, Hypothermia, Resample, Weka

I. INTRODUCTION

Data Mining is about elucidating the past and predicting the future by process of data analysis. It is a field which integrates the data, machine learning, database technology and artificial intelligence. Data mining is an application of automatically searching the huge reserve of data to uncover the patterns and trends that go beyond easy analysis. Data mining plays a significant role in the healthcare industry which enables the health systems to recognize the inefficiencies as well as practices to improve care by systematically using the data and analysis. The process has been developed by the health systems so that the patient's receive suitable care or treatment. Medical databases are ever-growing information in data mining field collected from hospitals about patients and their medical conditions. The dataset occur from the Irvine repository is the Post Operative Patient dataset. In which the patient shift from one unit to another after surgery and study of the complete picture of patient's condition to minimize the risk of medical errors and to offer the optimal patient care. The aim is to predict whether a patient will be sent to the general hospital floor (A) or prepared to go home (S) according to their health situation. Moreover, two patients were included in the training data sent to an Intensive care unit (I). [1]

The paper is arranged as: Section I discusses data mining in medical field. Similar works in the field of medical data mining are presented in Section II. Section III discusses the proposed method. Our experimental results with which dataset is mined are described in Section IV which is followed by a conclusion and future work in Section V.

II. DISCUSSION ON SIMILAR PAPERS

This section presents with an overview of various research papers contribution related to the techniques of data mining like Multi-layer perceptron, Decision tree, Random forest, Naïve Bayes, K-nearest neighbour for diagnosis in the medical field. Many researchers worked on different datasets in medical field, some of them have been briefly discussed as follows.

Authors predicted Quasi-Newton method performs well to predict the patient's post operative patient recovery area by comparing the seven algorithms to train the multi-layered neural network architecture. After that in the next level, Levenberg Marquardt performs better but both these methods are not suitable for large datasets. [3] Authors have compared three data mining prediction methods for breast cancer survivability by using 10-fold cross-validation. Decision tree (C5) is predicted as best with accuracy 93.6% on holdout sample, artificial neural network is second with accuracy of 91.2% and logistic regression is found to be third worst with 89.2% accuracy. [2] Authors predicted the most efficient model for the patients with lung cancer is Naïve Bayes than decision tree and neural network. [8] Authors has predicted that Support Vector Machine found to be best in prediction of cardiovascular disease in patients by comparing RIPPER, decision tree (C4.5), Artificial Neural Network, Support Vector Machine on basis of Accuracy, Sensitivity, True Positive Rate, Specificity, Error Rate, and False Positive Rate. [6] Authors explored the comparative study of the different classification techniques which involves the decision tree, fuzzy analogous concepts and Bayesian classification. Initially, the result unveils that training dataset is better to be used than use 10 cross fold. Consequently, by using weka, ID3 algorithm is predicted as the best algorithm. [7] Authors worked on unknown primary tumours where the classifier multiclass random forest is predicted for the classification of multiclass dataset which exhibit high accuracy than binary classifiers. To balance the dataset, SMOTE method is employed in order to improve the selected classifier results showing higher accuracy. [5] Author proposed that K-means method is used in medical database to deal with clustering. Some pre-processing technique is used to increase 81% accuracy and besides algorithm shows 94% accuracy for improvement of data. Consequently, new instances (with 700 records) used to show 97% of precise accuracy. [4]

III. PROPOSED METHODOLOGY

Classification is a data mining function which is a collection of records (training set) and each record consist of attributes(like Mean absolute error, ROC area, Root absolute error, Root mean square error, Kappa Statistics, Relative root square error, Time taken and Percentage value of correctly classified instances). Then find a model for attributes as a function of the values of other variables. Training set is used to build the model and test set is used to evaluate it. In this paper, *Fig. 1* shows the proposed work in which pre-processing method is used to remove the missing values by applying the unsupervised filter and stacked approach is applied on dataset which combine the two base classifiers: Decision Table, Naïve Bayes and a Meta classifier: J48 as decision tree. Output classifiers which perform best are used to make the predictions of instances on the test dataset.

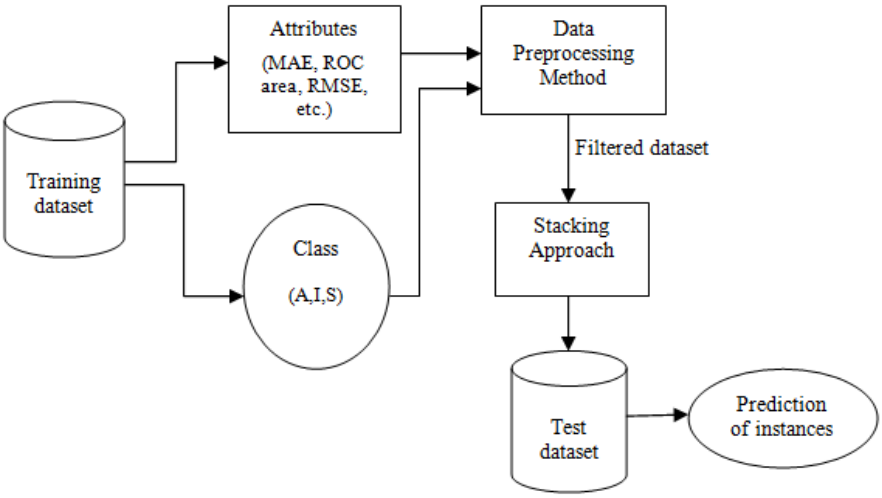

Fig. 1: Proposed Method

IV. ANALYSIS OF RESULTS

This section provides the information related to the dataset and the results obtained from the experiments in the form of graphical representation.

A. Data Source

Post-Operative Patient dataset were created by Sharon Summers and Linda Woolery and the donor is Jerzy W. Grzymala-Busse. The task of classification is to determine where the patients should be sent to next after post operative recovery area. Because hypothermia is main concern after the surgery, the attributes corresponds roughly to body temperature measurements was taken from UCI Machine Learning Repository. [1] *Table I* shows the summary of the dataset related to patient features.

TABLE 1: INFORMATION OF PATIENT FEATURES

ATTRIBUTES	LABEL	MISSING VALUES	UNIQUE
L-CORE	High, Low, Mid	No	0%
L-SURF	High, Low, Mid	No	0%
L-O2	Excellent, Good	No	0%
L-BP	High, Low, Mid	No	0%
SURF-STBL	Stable, Unstable	No	0%
CORE-STBL	Mod-stable, Stable, Unstable	No	1%
BP-STBL	Mod-stable, Stable, Unstable	No	0%
COMFORT	05, 07, 10, 15	3%	1%
DECISION ADM-DECS	A, I, S	No	0%

B. Pre-Processing

By using the unsupervised filter, the missing values are removed from "COMFORT" attribute through "ReplaceMissingValues" attribute in the training dataset. In *Fig 2* by using attribute selection method "InfoGain with Ranker Search Method", the attributes obtain rank according to their importance in the dataset as L-CORE is very important and essential attribute which gets first rank, after that L-SURF gets second rank and so on.


```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 9 decision):
        Information Gain Ranking Filter

Ranked attributes:
 0.26118    3 L-O2
 0.14899    8 COMFORT
 0.13398    7 BP-STBL
 0.09875    1 L-CORE
 0.0901     2 L-SURF
 0.04622    4 L-BP
 0.04451    6 CORE-STBL
 0.0049     5 SURF-STBL

Selected attributes: 3,8,7,1,2,4,6,5 : 8
```

Fig. 2: Ranking of Attributes through Ranker Search

To remove the biasness among the classes by resampling the dataset through "Resample" filter of supervised technique. *Fig. 3* shows "bias to uniform class" is "1.0" to make a uniform class distribution and "samplesizepercent" increases to "500.0".

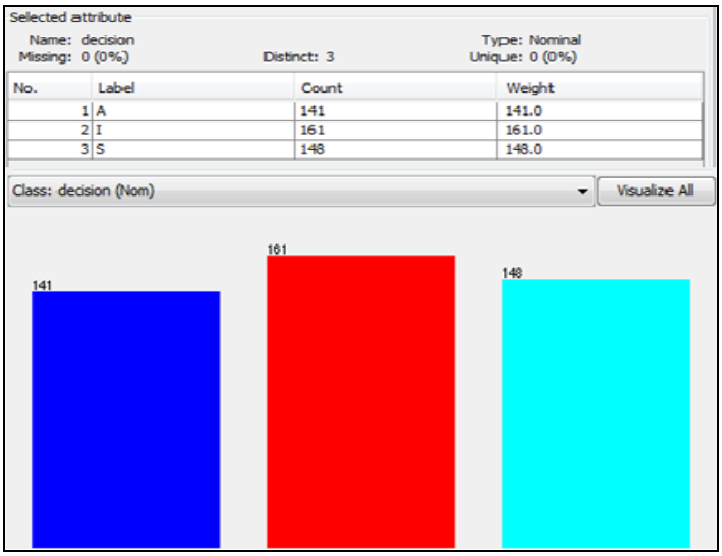

Fig. 3: Uniform Distribution of Classes

Stacked based technique having Naïve Bayes, Decision Table as base classifiers and J48 as Meta classifier provides better results with 94.67% accuracy. *TABLE II* shows the comparison among different classification techniques.

TABLE II: COMPARISON OF ALGORITHMS

ALGORITHM	**PRE-PROCESSING**	
	BEFORE	**AFTER**
Multilayer Perceptron	55%	90%
Naïve Bayes	66%	68%
J48	70%	92%
IBk	67%	92%
Random forest	62%	92%
Decision Table	68.8%	91.3%
Stacked Technique	62%	94.67%

Fig. 4 represents the graphical representation of comparison of before and after pre-processing of different classification techniques.

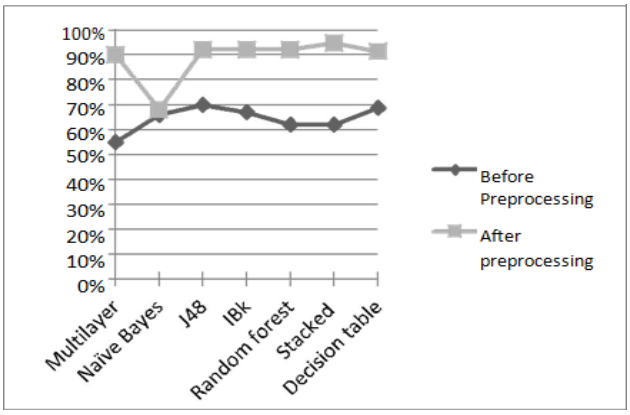

Fig. 4: Graphical representation of comparison of classification techniques

C. Testing

Testing is performed on the test dataset which is taken from .arff file consist of 13 instances having unknown classes. On the basis of selected stacked technique having 94.67% of maximum accuracy, the predicted values of each instance can be calculated which shows the predicted classes from discharge decision (A, I, S) on the unknown test dataset. *Fig. 5* shows the output predicted values of 13 instances of test dataset in WEKA.

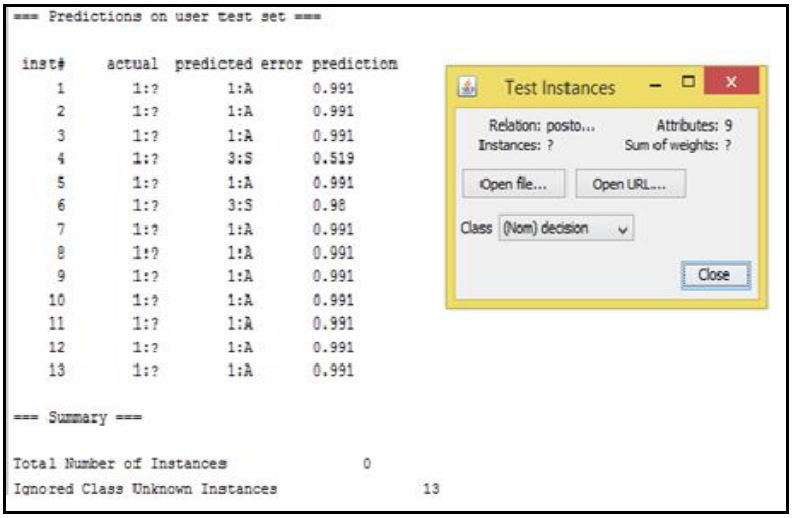

Fig. 5: Output predicted values of classes on test dataset

V. CONCLUSIONS

Testing has been performed on the training dataset in which stacked technique is predicted as best in which resampling has been done to eliminate the biasness among the classes and after pre-processing Class A (patient sent to general hospital floor) predicted with higher count in the test dataset. Class I and Class S can be better only if their count would be balanced in the training dataset before pre-processing. In future, variations among the algorithms can be done for better accuracy but according to our variations stacked technique is best which can work just like an expert system for the doctors to predict the classes of test dataset based on any training dataset. In future accuracy can be further enhanced by modifying or inserting any attribute which is necessary to measure after surgery against hypothermia and by applying the filters on the dataset.

ACKNOWLEDGMENT

This research paper is made feasible through the continuous support and helps from my parents and guide. They were prominent source of encouragement, enthusiasm, patience and great knowledge.

REFERENCES

[1] Dataset collected, [http://tunedit.org/repo/UCI/ postoperative-patient-data.arff] accessed 2015.

[2] D. Delen, G. Walker, A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods", Elsevier Science Publishers Ltd., Volume 34, Issue 2, Pages 113-127, June 2005.

[3] D. Shanthi, Dr. G. Sahoo, Dr. N. Saravanan, "Comparison of Neural Network Training Algorithms for the prediction of the patient's post-operative recovery area", Journal of Convergence Information Technology, Volume 4, Number 1, March 2009.

[4] Dr. B. M. Hussan "Data Mining based Prediction of Medical data using K-means algorithm", Basrah Journal of Science (A), Vol. 30(1), 46-56, 2012.

[5] M. Naib and A. Chhabra, "Predicting Primary Tumors using Multiclass Classifier Approach of Data Mining", International Journal of Computer Applications (0975 – 8887), Volume 96, No. 8, June 2014.

[6] M. Kumari, S. Godara, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", International Journal of Computer Science and Technology, Vol. 2, Issue 2, June 2011.

[7] S. R. Dash and S. Dehuri, "Comparative Study of Different Classification Techniques for Post Operative Patient Dataset", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 5, July 2013.

[8] V. Krishnaiah, Dr. G. Narsimha, Dr. N. Subhash Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", International Journal of Computer Science and Information Technologies, Vol. 4(1), 2013, 39 - 45.

