Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.8 No.S1, 2019, pp.36-37
© The Research Publication, www.trp.org.in
DOI: https://doi.org/10.51983/ajcst-2019.8.S1.1963

Study of Ensemble Classifier for Prediction in Health Care Data

S. Sathurthi[1], R. Kamalakannan[2] and T. Rameshkumar[3]
[1]Research Scholar, Pondicherry Engineering College, Puducherry, India
[2]Assistant Professor, Prist University, Puducherry, India
[3]Assistant Professor, Pondicherry Institute of Co-operative Management (PICM), Puducherry, India
E-Mail: sathurthisankar@gmail.com,kamalsmvec@gmail.com,trksolar09@gmail.com

(Received 4 December 2018; Accepted 24 January 2019; Available online 3 February 2019)

Abstract **–Electronic health record systems are adapted in a good deal of health care facility to improve the quality of patient care which is maintained electronically. Developing a disease prediction model for health care system can help us to overcome the problem of medical distress. In this study, we suggest ensemble technology and statistical methods to search through massive amounts of information, analyzing it to predict outcomes for individual patients. Using Weka tool, breast-cancer and diabetes medical datasets have experimented with ensemble classifier.**

Keywords: **Ensemble, Random Forest, Bagging and Boosting**

I. INTRODUCTION

Electronic Health Records are widely adopted in many healthcares in an attempt to improve the quality of patients care. Electronic health records (EHRs) that use structured data elements are documenting patient information using controlled vocabulary rather than narrative text. Health care data size is generally growing from day to day. With this large amount of data, the ability to extract useful knowledge hidden in these large amounts of data and to act on the knowledge is becoming increasingly important in today's competitive world.

The process of applying computer-based information system (CBIS), including new techniques, for discovering knowledge from data is called data mining. The process of machine learning is similar to that of data mining. Machine learning algorithms are often categorized as supervised learning and unsupervised learning.

Predictive modelling is a branch of clinical and business intelligence (C&BI) that is used to forecast the future health status of individuals and to classify patients by their current health risk. It can also be used in risk adjustment to compare the aggregate health risks of one physician's or one organization's patients to those of another doctor or healthcare entity. An ensemble of classifiers is combinations of multiple classifiers, referred as base classifiers. Ensembles usually achieve better performance than any of the single classifiers. Ensemble methods differ in the way they induce diversity between the base classifiers.

The most common approach is modifying the training set for each member of the ensemble. Ensemble methods play important role in predictive modelling.

II. METHODOLOGY

Ensemble learning is easily recognized approach used in supervised learning for prediction by combining various ensemble models [1].Ensemble methods enhance the accuracy and strength for building a classification model by combining a collection of base learners. The main goal of ensemble methodology is to scale back variance and bias. Ensemble methods used for classification are Boosting, Bagging and Random Forest.

A. Bagging

Bagging is one of the ensemble methods for improving the results of classification algorithms. This method was proposed by Leo Breiman and it is derived from the phrase "bootstrap aggregating" [2]. Bagging falls under the category of concurrent methodology. The main goal of bagging is to resolve the over fitting issues and the improvement of classifier accuracy.

B. Boosting

Boosting [3] is another kind of dependent ensemble methodology proposed by Freund and Schapire based on the learning in sequence [4].AdaBoost (Adaptive Boosting), which was first introduced by Freund and Schapire in 1996, could be a well-liked ensemble technique that improves the straightforward boosting rule via repetitive method [5].

C. Random Forest

Random forest aggregates the ideas of bagging and random subspace methodology. Random forest [6] achieves the internal estimation of error, strength, correlation and variable importance. The selection of single learning having high generalization accuracy is chosen from the group of base learner produced from the classifiers outputs.

D. Majority Voting

Majority voting [7] is one of the prominent weighting methods. Every base level classifier votes for one class label and therefore the final output class is chosen by receiving over half of the votes. If no one class labels receive over half of the votes, that they are rejected for the prediction model. Table 1 shows the comparative analysis of different healthcare datasets used in ensemble methods by different authors and their limitations.

TABLE I ANALYSIS OF VARIOUS DATASETS FOR PREDICTION MODEL

Author Name	Datasets	Limitations
Ping li *et al.*, method (2013)	Yeast Genbase	AdaBoostPDT performs worst since it does not consider the correlations among multiple labels.
Neeshajothi *et al.*, method (2015)	Diabetes dataset	Depends on feature and size of dataset between training and testing sets.Majority and the minority classifier are not balanced resulting prediction erroneous. Another limitation of healthcare data sets are the missing values.
Jiazhu *et al.*, method (2015)	Diabetes datasets	Multi classifier system perform worse when design is not proper
Jing zhao *et al.*, method (2015)	Clinical datasets	Extracting training data directly from EHR database
Jitendra *et al.*, method (2015)	Heart disease Datasets	System fail to identify risk factor data in few situation
Jose F.Diez-Pastor *et al.*, method (2015)	Ecoli Glass	For f-measure the basic Rotation Forest algorithm does not achieve a good position, although various combinations of Rotation forest with other ensembles still occupy top position.
Yanli *et al.*, method(2016)	Diabetes datasets	Accurate ways to measure the difference of data distributions among multiple participators under privacy constraint
John wes Solomon *et al.*, method(2016)	Blood pressure datasets	Erroranalysis uncovered that the model is more error prone when theabsolute value of actual SBP change is high, and that the model has a tendency to over-predict SBP values.
SurangaN. Kasthurirathne *et al.*, method(2016)	Cancer disease Datasets	The result of manual feature selection method depended on clinical expertise of the reviews and their familiarity with pathology report content

III. EXPERIMENT RESULT

TABLE II DESCRIPTION OF BREAST-CANCER AND DIABETES DATASETS

Dataset	No. of Attributes	No. of Instances	No. of Classes	Missing Values	Correctly classified instances	Incorrectly classified instances	Accuracy
BreastCancer	10	286	2	Yes	199	87	70%
Diabetes	9	768	2	No	582	186	75.78%

In this study, we have got performed random forest classifier on Breast-cancer and Diabetes datasets with help of WEKA tool. In Breast-cancer and Diabetes datasets has 286 and 768 instances with two binary classes. We have a tendency to found the missing values, correctly and incorrectly classified instances of every dataset in this experiment. Based on evaluation metrics, the performance of the classifier may be evaluated. The accuracy is one of the distinguished evaluation metrics that is calculated for each dataset. The description of each dataset is given in the following Table II.

IV. CONCLUSION

In this study we had a brief discussion on ensemble methods for prediction model of electronic healthcare data and implement the random forest algorithm with breast cancer and diabetes datasets. The training data of each dataset is tested with cross-validation method and accuracy can be calculated by correctly classified instances in the dataset. This research work focuses on the accuracy of the predicted model using an ensemble method.

REFERENCES

[1] T. G. Dietterich, “Ensemble methods in machine learning”, *In: Proceedings of Multiple Classifier System, Springer,* Vol. 18, pp. 1–15, 2000

[2] L. Breiman: “Bagging predictors”, *Mach. Learn.,* Vol. 24, No. 2, pp.123–140, 1996.

[3] Thomas G. Dietterich, “An Experimental Comparison of Three Methods for Constructing of Decision Trees: Bagging, Boosting, and Randomization”, *Machine Learning,* Vol.40, pp. 139–157, 2000.

[4] Y. Freund, and R.E. Schapire, “Decision-theoretic generalization of on-line learning and an application to boosting”, *J. Computer and System Sciences* Vol. 55, No. 1, pp.119–139, 1997.

[5] LiorRokach, “Ensemble Methods for Classifiers”, *Data Mining and Knowledge Discovery Handbook,* pp. 957-980.

[6] L. Breiman. “Random forests”, *Machine Learning,* Vol. 45, No. 1, pp.5–32, 2001.

[7] D. Wolpert, “Stacked generalization”, *Neural Networks,* Vol. 5, No. 2, pp. 241–260, 1992.