INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue2)

Healthcare Prediction Analysis in Big Data using Random Forest Classifier

P. Subhapriya
Assistant Professor
Department of Computer Science and Engineering
Sri Manakula Vinayagar Engineering College,
Puducherry, India
subhaji.sach@gmail.com

R. Sujatha
UG students
Department of Computer Science and Engineering
Sri Manakula Vinayagar Engineering College,
Puducherry, India
sujatharaju27@gmail.com

K. Meghana
UG students
Department of Computer Science and Engineering
Sri Manakula Vinayagar Engineering College,
Puducherry, India
meghanakrishnan14@gmail.com

Abstract: An infrastructure build in the big data platform is reliable to challenge the commercial and not- commercial IT development communities of data streams in high dimensional data cluster modeling. The knowledge discovery in database (KDD) is alarmed with the development of methods and techniques for making use of data. The data size is generally growing from day to day. One of the most important steps of the KDD is the data mining which is the ability to extract useful knowledge hidden in this large amount of data. Both the data mining and healthcare industry have emerged some of reliable early detection systems and other various healthcare related systems from the clinical and diagnosis data. In this paper propose the enhanced data mining algorithm for healthcare application. It consists of three steps they are anomaly detection, clustering, and classification. In this classification algorithm use the random forest algorithm for accurately predict the patient result from a large amount of data. Finally, our experimental result shows our proposed method can achieve more accuracy result.

Keywords: Random Forest Algorithm, Knowledge Discovery in Data (KDD), Big Data, Data Mining.

I. INTRODUCTION

Health care systems are highly complex, fragmented and use multiple information technology systems. With vendors incorporating different standards for similar or same systems, it is little wonder that all-around inefficiency, waste, and errors in healthcare information and delivery management are all too commonplace an occurrence. Consequently, a patient's health records often get trapped in silos of legacy systems, unable to be shared with members of the healthcare community. These are some of the several motivations driving an effort to encourage standardization, integration and electronic information exchange amongst the various healthcare providers.

The study termed as Developmental Origins of Health and Diseases or DOHAD has successfully proven the importance of developmental records of individuals in predicting and/or explaining the diseases that a person is suffering from. In the current largely paper-based health records world, invaluable data is more often than not unavailable at the right time in the hands of the clinical care providers to permit better care. This is largely due to the inefficiencies inherent in the paper-based system. In an electronic world, it is very much possible, provided certain important steps are taken beforehand, to ensure the availability of the right information at the right time. Supervised learning is the machine learning task of interfering a function from labeled training data. We are using supervised learning for training set in this project.

The ensemble of classifiers is combinations of multiple classifiers, referred as base classifiers. Ensembles usually achieve better performance than any of the single classifiers. In order to build a good base classifiers, also the base classifiers must be diverse, this means that for the same instance, the base classifiers return different outputs and their errors should be in different instances.

In this project, we develop a hybrid predictive model for Electronic Health Records (EHR) using ensemble method and calculate classifier accuracy by analyzing the limitations of the previous developments and a few advancements are made to the project.

II.PROBLEM SPECIFICATIONS

- Medical datasets are often not balanced in their class labels.
- Most of the existing classification methods tend to perform poorly on a dataset which is extremely imbalanced.
- An increasing number of applications deployed over the cloud operate on datasets which are large and complex that it becomes difficult to gather, store, analyze and visualize. So there arises a scalability issue.
- In this project imbalance issue and scalability issue is overcome.

III.LIMITATIONS

- Combining models, in case trees will only be beneficial if the individual models are accurate and different from each other.
- The data mining methods accuracy varies depending on the features of the data sets and the size of data set between the training and testing sets.
- The common characteristics of the healthcare data sets are highly imbalanced data sets, whereby the majority and the minority classifier are not balanced resulting prediction erroneous when run by the classifiers.
- Other characteristics of healthcare data sets are the missing values.
- Using only structured data in healthcare datasets

IV. SCOPE

- Medical data analysis
- Bioinformatics
- Drug event detection
- Clinical investigation

V. EXISTING SYSTEM

Our existing system uses breast cancer datasets which have two classes' recurrence events and no recurrence events. Preprocess the data and then classifying the data using decision stump which is to be used as a base classifier for AdaBoost algorithm with number of iteration is set to 2 and weight threshold for weight pruning is set to 10 AdaBoost.M1 algorithm is used, which use the base classifier Decision Stump(AdaBoost_DS) and reweighting, the number of iterations is set on 10, and weight threshold for weight pruning is set on 100.Comparing the correctly classified instance and accuracy of classification, AdaBoost implementation of decision stump improves accuracy.**AdaBoost,** short for "Adaptive Boosting", is a machine learning meta-algorithm, It can be used in conjunction with many other types of learning algorithms to improve their performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier.

VI. PROPOSED SYSTEM

In this project, we suggest OVI Random Forest as classifier ensemble that can incorporate different base classifiers into classifier ensembles models for classification problems. This project suggests that the impact of using different base classifiers on classification accuracy of Random Forest classifier ensemble. Classifier ensembles with five base classifier have used on five medical data sets. These results evaluated and compared choosing a different type of decision tree algorithms for the base classifier. The reliability of classification for most of datasets and classifier ensembles is increased when we select the appropriate j48 random forest classifier achieves the minimum time required to build models.

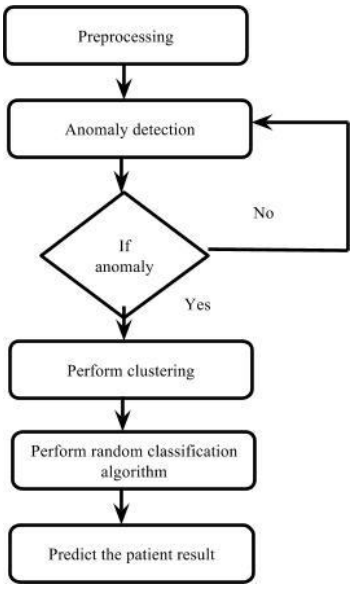

Fig., 1.1 Flow Chart

VII.SYSTEM REQUIREMENTS

Minimum Hardware Requirements

- System: Pentium IV 2.4 GHz.
- Hard Disk : 40 GB.
- Monitor : 15 inch VGA Color.
- Mouse : Logitech Mouse.
- Ram : 512 MB
- Keyboard : Standard Keyboard

Minimum Software Requirements

- Operating System: Windows XP.
- Platform : Java TECHNOLOGY
- Tool :NetBeans6.9.1,Hadoop2.3,weka
- Front End : Jdk 1.7

Tools Used

- Java ML Library
- Weka

VIII. ADVANTAGES

- It is simple to understand and interpret and able to handle both numerical and categorical data, which requires little data preparation, for possible to validate a model using statistical tests, performs well with large datasets.
- It is robust, which means that performs well even if its assumptions are somewhat violated by the true model from which the data were generated.

CONCLUSION

The data mining has played an important role in the healthcare industry, especially in predicting various types of diseases. The diagnosis is widely being used in predicting diseases; they are extensively used in medical diagnosing. In conclusion, there is no one data mining method to resolve the issues in the healthcare data sets. In order to obtain the highest accuracy among classifiers which is important in medical diagnosing with the characteristics of data being taken care, we need to design a hybrid model which could resolve the mentioned issues. In future, more work will be done in this field so to improvise the treatments and the lifetime of the patient by properly maintaining and analyzing the health sector data for directions is to enhance the predictions using hybrid models.

REFERENCE

[1] Yan Li, Changx in Bai, Chandan K.Reddy, A distributed
ensemble approach for mining health care data under privacy constraints, Information Sciences 330 (2016) 245–259

[2] Muhammad A.Hasan a,n, Vijay S.Chauhan b, Sridhar Krishnan, Beat-to-beat T-wave alternate detection using the Ensemble Empirical Mode Decomposition method, Computers in Biology and Medicine 77 (2016) 1–8

[3] Ping Li, Hong Li, Min Wub, Multi-label ensemble based on variable pairwise constraint projection, Information Sciences 222 (2013) 269–281

[4] F. Di Maio, J. Hu, P. Tse, M. Pecht, K. Tsui, E. Zio, Ensemble-approaches for clustering health status of oil sand pumps, Expert Systems with Applications 39 (2012) 4847–4859

[5] Wei-Liang Tay, Chee-Kong Chui, Sim-Heng Ong, Alvin Choong-Meng Ng, Ensemble-based regression analysis of multimodal medical data for osteopenia diagnosis, Expert Systems with Applications 40 (2013) 811–819

[6] Elders Quan Liua,1, Xingran Cuia,b,1, Yuan-Chao Chouc, Maysam F. Abbodd, Jinn Line,Jiann-Shing Shieh, Ensemble artificial neural networks applied to predict the key risk factors of hip bone fracture for elders, Biomedical Signal Processing and Control 21 (2015) 146–156

[7] Hristijan Gjoreski, Bostjan Kaluza, Matjaz Gams, Radoje Mili, Mitja Lustrek, Context-based ensemble method for human energy expenditure Estimation, Applied Soft Computing 37 (2015) 960–970

[8] Cagatay Catal, Selin Tufekci, Elif Pirmit, Guner Kocabag, On the use of ensemble of classifiers for accelerometer-based activity Recognition, Applied Soft Computing 37 (2015) 1018–1022

[9] Nikunj C. Oza, Kagan Tumer, Classifier ensembles: Select real-world applications, Information Fusion 9 (2008) 4–20

[10] Robi Polikar, Apostolos Topalis , Devi Parikh, Deborah Green, Jennifer Frymiare, John Kounios, Christopher M. Clark, An ensemble-based data fusion approach for early diagnosis of Alzheimer's disease, Information Fusion 9 (2008) 83–95

