ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 3, Issue 6, June 2018

Analysis of Different Data Mining Classifiers for Predicting Angiographic Disease Status

[1] Ranjani Murali
[1] Computer Engineering Department Sardar Vallabhbhai National Institute Of Technology Surat, India

Abstract: ***The healthcare industry's data generated has increased exponentially due to reforms in technology, use of IOT and intensive patient care. This necessitates use of data mining technology for efficient processing and decision making. The early detection and prediction of curable diseases in healthcare can be done via use of this intelligent tool. This paper compares use of different classifiers in data mining for predicting Angiographic disease status.***

Index Terms— ***Data Mining, Health informatics, Classifiers, Healthcare, Naïve bayes, Kstar, Random tree, Random Forest, Logistic model tree, AdaBoost, J48, ZeroR, Kmeans.***

I. INTRODUCTION

The recent developments seen in Healthcare industry from the increase in technological advances, Electronic health records (EHR), Personal health monitoring devices with IOT (Internet Of Things) and ubiquitous monitoring by household devices installed via IOT has exponentially increased the data generated and augmented the complexity of its processing.
The potential of using Data mining and Artificial intelligence in healthcare is immense from helping healthcare insurers detect fraud and abuse, helping physicians identify effective treatments and best practices to helping patients to receive more effective and affordable healthcare services. It can serve a ubiquitous role in the healthcare industry as it offers benefits to care providers, patients, healthcare organizations, researchers, and insurers. In healthcare, data mining has been proven to be expedient in areas such as, customer relationship management, predictive medicine, detection of fraud and abuse, management of healthcare and measuring the efficacy of certain treatments. Data mining can not only be used for analysis and cause effect relation establishment in medical data but also be used to predict and prevent diseases at the germinal stage.

One of the most important applications of data mining is the early detection and prediction of curable diseases in healthcare. The use of readily available data of patients from their birth can be used to predict and analyze possibility of curable disease via data mining. The detection at an early stage of certain disease can increase the chances of cure considerably. The increase in personal monitoring devices, android applications to IOT have increased the avenue and features of data sources and have enable efficient monitoring of patients.
Heart or Cardiovascular diseases are one of the major areas of healthcare where timely detection and treatment can save the patient's life. Cardiovascular diseases include a wide series of conditions that affect the heart and the blood vessels and the modus in which blood is pumped and circulated through the body. Cardiovascular disease can at advance stages result in severe illness, disability, and also death. Chest pains arise due to inadequate blood supply. This if done in an early stage can be cured by stunting the arteries or via heart surgery.

Detection of angiographic disease needs to be both accurate and efficient. This arises the need to use state of the art, ubiquitous, pervasive detection and analysis system which should be integrated with data sources alternating from personal health monitoring devices, regular hospital reports to home applications monitoring health data. The information from these various sources can be collected remotely and stored in the hospital data warehouse and analysis of the patient can be completed by the doctor with more efficiency. The system has automated monitoring and early warning system as part of its protocol to detect the known cardiac patient's condition deterioration. This early detection of either the increase in blood pressure and anomalous increase of heart rate and other indicators can

automatically warn the patient beforehand and the concerned cardiac specialists about the status of the patient. Early cautionary systems can be used to actively track and help the patient before severe symptoms.

The detection and analysis system has data mining and classifiers as its decision constructing component which based on previous data establishes rules and patterns for prediction of angiographic disease status. The training data set fed into the system enables it to efficiently predict and identify the anomalous state of patient health and send warning reports as well as any emergency services necessary for the patient.

II. LITERATURE SURVEY

Several research work has been endeavored in the interdisciplinary area of healthcare data mining. Applications in this area include usage of Artificial Neural networks, Fuzzy systems, Associative rule mining, Markov models, Decision tree, Inductive mining, Sequence and Pattern discovery and Genetic algorithm based healthcare information systems in the process mining survey in [2] [9] [5].

Several synthetic intelligence techniques like SVM in [6] used for diabetes detection, Decision tree and neural network in [7] for heart disease status and use of RBF neural networks to investigate for survivability of breast cancer have been pursued.
The exponential increase of healthcare data needs use of big data management techniques where the electronic health data, medical imagining data, unstructured clinical notes and the genetic data is analyzed as explained in [4][15]. This research work discusses the use of Hadoop architecture to explore the clinical health care data by using big data analytics to produce decision support for the medical practitioners and increases the efficiency of the diagnosis. It also describes the extensive use of wearable sensors and the data potential for diagnosis.
The recent increase in wearable devices and health monitoring IOT based devices data are used to analyze patient data in [3] where a scalable cloud based architecture is used to maintain the patient data and stochastic gradient descent algorithm is used in the logistic regression to develop the scalable diagnosis model. The sensor data of blood sugar level, heart-rate and blood pressure data collected from the sensors is remotely processed and used to monitor the patient's angiographic status and performs detection with an accuracy of 81.99%.

III. DATA MINING CLASSIFIERS AND BIOINFORMATICS

Data mining
The amalgamation of computer science and statistics is called data science. The discovery of hidden patterns in data, prediction of future data based on the past trends and discovery of similarity of data all are some applications of data mining. The use of data mining in Financial and banking sectors has resulted due to the early legacy systems using statistical techniques. Combination of computer science with statistics has increased the system's efficiency and capacity. Several data mining techniques have been analyzed in this paper. A brief description is given below.
Naïve Bayes: It is a simple probabilistic classifier based on applying Bayes' probability theorem with strong independent assumptions and is particularly suited when the dimensionality of the inputs is high i.e. more parameters to be assessed and classes in larger number. It assumes independence of the parameters supplied. It predicts categorical labels (or discrete values) in a two iterations. In the first iteration, a model is built describing a predetermined set of data classes or concepts. The model is constructed by analyzing database rows described. Each row is assumed to belong to a single class called the class label attribute. In the second iteration, the model is used for classification.

Random Forest and Random Tree: They are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multiple horde of decision trees at training time and output the classification or mean prediction (regression) of the individual trees.
ZeroR and OneR: OneR (One Rule) is classification algorithm that generates one rule for each class in the data, then selects the rule with the smallest total error as its singular rule. An initial frequency table is constructed by this classifier to assess the errors. It has been shown that OneR produces rules only slightly less accurate than state-of-

the-art classification algorithms while producing rules that are simple to interpret. Whereas ZeroR use mean or mode prediction as its basic mechanism of classification where no single rule exits for its class determination like OneR.
Kstar: It is an instance-based classifier that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. It differs from other instance-based learners in that it uses an entropy-based distance function.

J48: It is an implementation of ID3 (Iterative Dichotomiser 3) algorithm with use of information gain, a concept that measures the amount of information contained in a set of data. It gives the idea of importance of an attribute in a dataset. This parameter is analyzed to classify given set of data. The rules applied here classify instances based on the information gain parameter and the class labels are then predicted after training the classifier.
Health care informatics and Data mining
The term healthcare informatics refers to information engineering applied to the field of health care and use of patient healthcare information. It is a multidisciplinary field that uses health information technology to improve health care via higher quality, higher proficiency, and new opportunities by either providing decision support, cost efficiency or through data management for admistration.The types of data in healthcare include the following:
Provider data: The data and knowledge from medical experts with both tacit and explicit knowledge is provider or expert data. Doctors and expert's pre-requisite knowledge in the field to identify typical medical diagnosis or details from various available sources is classified as part of provider data. The Years of experience in medical diagnosis used by the medical practitioner to take better decisions.
Patient data: Implicit knowledge developed from the patients and it is considered "health status". The data from various sources like personal wearables, health monitoring devices and observations clinical unstructured data become classified in this category.

Organizational data: This data is also a vital role in patient treatments and analysis for preventative maintenance and illnesses. This data is developed from textual and electronic materials, medical analytic systems, past patient data and other resources.
The most familiar big data sources in medical environments include Electronic Health Record (EHR), Medical Imaging Data, Unstructured Clinical Notes and Genetic Data [4]. The potential to analyze healthcare data is immense and can not only increase the efficiency of the healthcare system in general but also provide intensive care via remote monitoring. The proposed work here describes an amalgamated system of healthcare with ubiquitous monitoring of patients and complete integration of the hospital database. The integrated system approach can both decrease the complexity of patient monitoring for the medical professionals and simultaneously provide intensive care to patient.

IV. FRAMEWORK

The integrated Healthcare data monitoring systems has 3 components which are Data Input sources, Database, Services. Each of the three components play an imperative role in the functionality of the system. Figure 1 represents an integrated system with the amalgamation of all the three components.
Data Input Sorces
The incorporation of Personal monitoring devices, Healthcare Wearable (IOT), Electronic health records, Medical Imaging Data, Unstructured Clinical Notes and Genetic Data are the data input sources. These comprise of a ubiquitous and consistent information sources for patient health monitoring. The data from these sources is collected and the hospital data base stores and processes this data continuously to detect any deterioration in the patient health status. The use of wearable and personal health care data increases the efficiency of monitoring and results in prompt detection of any emergency scenario for the patient

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 3, Issue 6, June 2018

Fig. 1.Hospital Data monitoring system

Data Base

The Data from the ubiquitous data sources are collected and processed in this module of the entire system. The data from these sources are stored and processed by the big data analytics system. The database is integrated with a cloud server and a data mining and classifier module. This data mining module is trained with past patient data and genetic history of the patients to enable it to recognize the data patterns corresponding to angiographic disease status. The continuous data from wearable sensors, personal health monitoring devices and Household monitoring devices are fed into the data base and are used to monitor intensive care patients with severe health conditions. The real-time processing is done by the data analysis system and the result data are again stored in the database. In case of a detecting an angiographic condition the system dispatches warning alerts to the medical personnel and initiates the request for emergency services on behalf of the patient. The immediate response of the services help in immediate recuperation and restoration of the patient under question.

Services

These refer to emergency services provided by automatic dispatch via the monitoring system. These also include preemptive diagnosis and result production. An early warning system is inbuilt which warns the patient of deteriorating condition in advance for recuperative action and also sends the monitored status to the doctors.

The data mining module structure and accuracy analysis has been explored in this paper. Different classifiers were used to test the accuracy of prediction and the results are described in the next section.

V. RESULTS

The data mining module was tested with 8 different classifiers for accuracy and system efficiency in detection of angiographic disease status with instances taken from 1277 patients and attributes for classification are 14.

ACCURAY ANALYSIS TABLE

Classifier name	Accuracy analysis parameters				
	Cross-folds	*Correctly classified instances (%)*	*Mean absolute error*	*Root mean square error*	*Kappa statistic*
AdaBoostM1	10	58.2616	0.3062	0.3838	0
	1000	58.2616	0.3062	0.3838	0
J48	50	81.8324	0.0807	0.2202	0.6924
	100	80.971	0.0846	0.2272	0.6774
Kstar	10	95.6147	0.0185	0.1136	0.928
	100	98.9037	0.005	0.052	0.9821
	1000	99.2169	0.004	0.0445	0.9872
Logistic Model Tree	10	94.5184	0.0333	0.1486	0.9099
	100	98.1989	0.0194	0.0938	0.9705
	1000	98.0423	0.0201	0.094	0.968
Naïve Bayes	10	61.9421	0.1655	0.3086	0.3644
	100	61.9421	0.1656	0.3083	0.3657
	1000	62.2553	0.1659	0.3087	0.3708
Random Forest	50	98.982	0.0653	0.1224	0.9833
	100	98.982	0.0649	0.1215	0.9833
Random Tree	10	95.5364	0.0279	0.135	0.9264
	50	98.2772	0.0197	0.0934	0.9717
	100	98.3555	0.0161	0.0841	0.9731
	1000	97.964	0.0192	0.0935	0.9666
ZeroR	10	58.2616	0.2446	0.3494	0
	100	58.2616	0.2445	0.3495	0
	1000	58.2616	0.2447	0.3497	0

The categories for classification angiographic status of chest pain types are typical angina, atypical angina, non-angina pain, asymptomatic. The data from 1277 patients included 14 parameters to help identify the pain category which includes basic demographic and genetic information, Cholesterol, blood-sugar levels, blood pressure levels, personal habits and other data.

The 8 classifier were analyzed using Weka 3.8.2 data mining tool. The parameters for accuracy analysis were classification accuracy percentage, mean absolute error, root mean square error and Kappa statistic.

Percentage Accuracy (P.A.)

The percentage accuracy of the classifiers is calculated by:

$$P.A. = \frac{Correclty\ classified\ instances}{Total\ instances} \quad (1)$$

The percentage of accurate classification is highest for Kstar, LMT, Random Forest and Random tree based classifiers. The classifiers J48 and Naïve Bayes show moderate performance with Adaboost and ZeroR showing the worst accuracy in classifying angiographic disease status. The low accuracy of ZeroR is due to the usage of mean and mode as its main logical module for classification. The categorical and tree based classifiers perform well due to the number of classification categories being moderate and the parameters of input being 14 which are three fold compared to the number of class instances which is 4. The highest accuracy being 99.2169% by Kstar and 98.982% by Random Forest classifier. Figure 2 shows the accuracy values along with the cross fold used to train the data.

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 3, Issue 6, June 2018

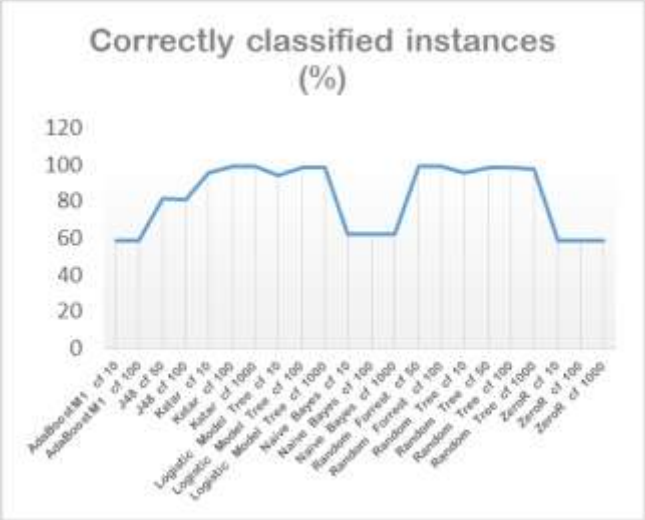

Fig. 1. Percentage accuracy of classification

Mean absolute error

The analysis based on mean absolute error also yields KStar, LMT, Random Tree and Random forest as the best performers with least error. The classifiers Naïve Bayes and J48 yield moderate results with worst performs being Adaboost and ZeroR. The number of cross folds i.e. the number of divisions for training and testing data values is observed to affect the individual classifier. The increase in the cross folds yields initial better results with lesser mean absolute error but after a certain threshold the error starts to increase yielding poorer results on increasing the number of cross fold. This is due to the over fitting problem of classifiers where the classifier starts to increase rules for classification by not being able generalize the rule set. The classifier instead assigns rules to categorize only the present data set instead of generating a pattern recognition mechanism for future data.

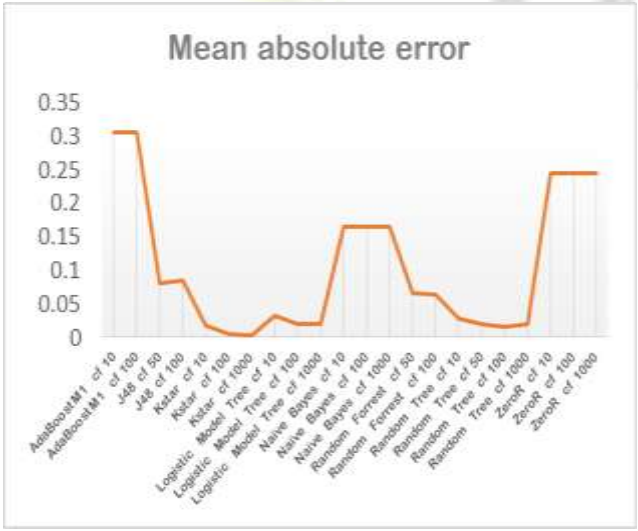

Fig. 1. Mean absolute error of classifiers

Root *Mean square error*

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 3, Issue 6, June 2018

The analysis based on root mean square yields results identical to mean absolute error where the most accurate classifier in Kstar classifier but is followed by Logical Model tree. The worst results are yielded by Adaboost due to conjunction of several weak classifiers being included together to reduce overall performance. This classifier perform poorly due to its dependence on the component classifier's ability and accuracy.

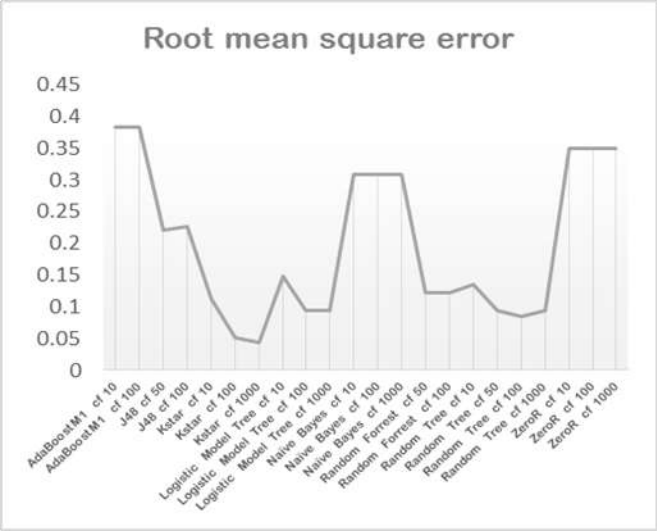

Fig. 1. Root mean square error of classifiers

Kappa Statistic

This parameter has a similar methodology like Percentage accuracy but it also takes into consideration the possibility of the agreement of classification occurring by chance. This analysis yields a slightly smaller figure than accuracy percentage value. The classifiers Kstar and Random forest yield the highest accuracy followed by Logical Model tree and Random tree based classifiers. The increase in number of cross folds has a similar effect on Kappa statistic parameter. The initial increase of cross folds increases the Kappa statistic but after a threshold there is a decrease in the parameter due to increase in cross folds. This classifier has a steep decrease for weaker classifiers as compared to the results based on the other parameters of assessment. The accuracy of classifier decreases below 60% then a steep decrease to 0 is seen in the Kappa statistic parameter effectively eliminating the weaker classifiers.

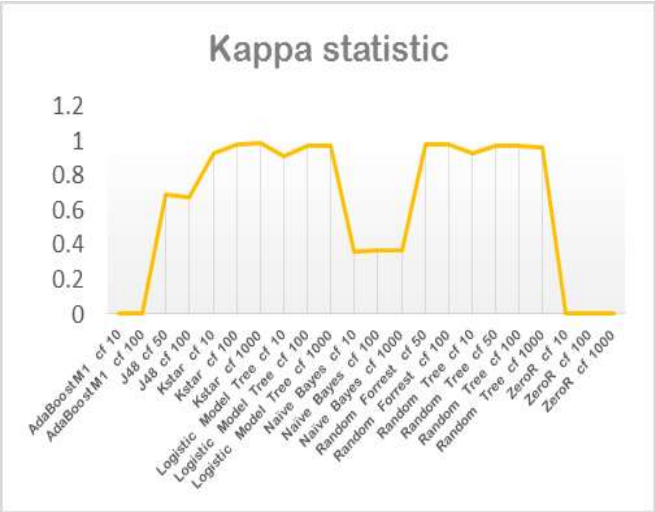

Fig. 2. Kappa Statistic

VI. CONCLUSION

The increase in data sources of the healthcare system with use of personal monitoring devices, wearables, and IOT has generated the need for synthetic intelligence systems. The intensive care necessary for patients and a decision support mechanism for medical practitioners can be supported via data mining. An early detection of disease status via use of synthetic intelligence can increase the chances of cure. This necessitates the use of a data mining system in health care systems for better and efficient care.

A system with health care data analysis was discussed and an angiographic disease status analysis module was proposed in this paper. Different classifiers where analyzed for the system under consideration. A dataset of 1277 patients and attributes for classification of 14 types were considered for training the classifiers. Kstar with percentage accuracy 99.2169% and Random Forest with percentage accuracy 98.982% accuracy yielded results based on comparison of different parameters of classification percentage accuracy, Kappa statistic, mean absolute error and root mean square error.

VII. ACKNOWLEDGMENT

The author is indebted to Dr. Rupa G Mehta Head of Department and Dr. Dipti P Rana Assistant professor, Computer Engineering Department SVNIT Surat for their guidance and support.

REFERENCES

[1] B.A.Venkatesh., et.al, "Cardiovascular event prediction by machine learning: the multi-ethnic studyof atherosclerosis," Circulation Research vol. 121, Issue 9, pp 1092-1101, October 2017.
[2] E. Rojas ,J. M. Gama , M. Sepúlveda and D. Capurro, "Process mining in healthcare: A literature review, " Journal of Biomedical Informatics vol 61, pp, 224–236, April 2016.
[3] G. Manogaran and D. Lopez, "Health data analytics using scalable logistic regression with stochastic gradient descent ," International. J. Advanced Intelligence Paradigms, vol. 10, pp. 1-2, 2018.
[4] G. Manogaran et.al., "Big Data Knowledge System in Healthcare ," Springer International Publishing, DOI 10.1007/978-3-319-49736-5_7.
[5] H. C. Koh and G. Tan, "Data mining applications in healthcare," Journal of Healthcare Information Management, vol. 19, no. 2, pp.64-72, 2005
[6] I. Kavakiotis et.al, "Machine Learning and Data Mining Methods in Diabetes Research," Elsvier publications, Computational and Structural Biotechnology Journal,vol. 15, pp.104–116, January 2017.
[7] K. Srinivas, B. K. Rani and A. Govrdhan, "Applications of Data Mining techniques in healthcare and prediction of heart attacks," International Journal on Computer Science and Engineering, vol. 02, no. 02,pp. 250-255, 2010.
[8] P. B. Jensen1, L. J. Jensen and S. Brunak, "Mining electronic health records towards better research applications and clinical care," Macmillan Publishers, vol.13, pp. 395-405 , June 2012
[9] P. Lucas. "Bayesian analysis, pattern analysis and Data Mining in health care," Current Opinion in Critical Care, vol.10, pp.399-403, 2004.
[10]
[11] R. Detrano et.el. "International application of a new probability algorithm for the diagnosis of coronary artery disease," American Journal of Cardiology, vol. 64, pp.304-310, 1989.
[12] S. Letchmunan, Z. Mansor, N. L.W.Yan , L. K. Meng, and N. F. I. Tahir, "Predictive analytic in health care using case-based reasoning," Proc. of the 6th International Conference on Computing and Informatics, pp. 8-5, April 2017.
[13] V. Chaurasia1 and S. Pal, "Data mining techniques: to predict and resolve breast cancer survivability," International Journal of Computer Science and Mobile Computing, vol. 3, no.1, pp. 10-22, January 2014.
[14] V. Chaurasia1 and S. Pal, "Data mining approach to detect heart dieases," International Journal of Advanced Computer Science and Information Technology, vol. 2, no. 4, pp. 56-66, 2013.
[15] Y. Wang and N. Hajli, "Exploring the path to big data analytics success in healthcare," Elsevier publivcations, Journal of Business Research, DOI: 10.1016/j.jbusres.2016.08.002, 2016.
[16] Y. Zhang, M. Qiu, C. W. Tsai, M. M. Hassan and A. Alamri, "Health-CPS: healthcare cyber-physical system assisted by cloud and big data," IEEE systems journal, 2015.
[17] http://archive.ics.uci.edu/ml/index.php