International Journal of Advanced Research in Big Data Management System
Vol. 3, No. 2 (2019), pp.21-26
http://dx.doi.org/10.21742/ ijarbms.2019.3.2.04

A survey on Machine Learning Classifiers and Big data for Accurate and Reliable Heart Disease Pre-diagnosis

Srikanth Meda [1]

[1]*Associate Professor, Dept. of CSE, R.V.R.& J.C. College of Engineering, Guntur.*
[1]*medasrikanth@gmail.com*

Abstract

Since a decade, emergence of interdisciplinary computer technologies changed the pace of medical diagnosis systems by insisting up-to-date intelligence and supervision. These intellectual systems predict the future health problems by processing the current health information of patients, which helps in prevention of diseases rather than cure. Although the medical diagnosis systems are adequate intelligent in disease diagnosis, but they are still suffering in pre-diagnosis of diseases due to the complexity in processing of huge medical datasets. Recently introduced Data Mining techniques with Big Data processing environment expanded the horizons of medical diagnosis systems to process the high velocity medical data sets to diagnose the occurrence of diseases early. Today's medical diagnosis systems, which are utilizing different data mining techniques like Decision Trees (DT), Support Vector Machines (SVM), Naïve Bayes (NB), Fuzzy Logics and K-Nearest Neighbor (KNN), are suffering from uncertainty, imprecision and complexity in processing. In this paper we are proposing a cross reference methodology to improvise the reliability and precision of diagnostic results and utilizing big data tools to diminish the complexity in processing huge sets of medical data. Most popular data mining techniques, which are participating in medical data processing with high accuracy are selected and cross referenced by our proposed framework to overcome uncertainty and imprecision. In order to process the high velocity medical datasets with several data mining techniques, this frame work outsources the data processing business to Apache Hadoop environment. Experiments on Cleveland medical dataset proved that the proposed cross reference methodology framework recorded high precision, recall and accuracy in results than its counterparts.

Keywords: *Data Mining, Machine Learning classifiers, Decision making systems, Disease diagnosis.*

1. INTRODUCTION

With the advent of advanced computer technologies intervention into medical diagnostic system, today most of them are being identified and treated in a better manner. Recently disease pre-diagnosis concept attracted the research scholars [1], which help in prediction of disease occurrence chances early by mining patient's health vault using data mining techniques. Pre-diagnosis happens by considering the present health conditions, past medical history and disease symptoms in an integrated manner, to support the popular medical statement "prevention is better than cure".

Article history:
Received (May 22, 2019), Review Result (July 3, 2019), Accepted (October 10, 2019)

Print ISSN: 2208-1674, eISSN: 2208-1682 IJARBMS
Copyright ⓒ 2019 GV Press

Some former researches[1][2] concentrated on disease pre-diagnosis concept and extended this concept development to their maximum possible level. In 2015,Randa. et al. [3]utilized the fast decision tree and C4.5 trees to diagnosis the health information from patient's medical datasets. In her work, Indira S. Dessai [4] offered a Probabilistic Neural Network (PNN) on K-means clustered heart disease database with 13 attributes, to extend the power of diagnostic system for classification and prediction of diseases. In 2010, AliAdeli et al.[5] designed a fuzzy expert system with the help of fuzzy logic techniques, to identify the coronary heart disease occurrence in early.

According to the WHO (World Health Organization) survey report, sudden attack of coronary diseases leads to human death in most of the cases and heart disease is recording high death ratio when compared with other diseases. Early detection of heart disease occurrence becomes very difficult due to many factors affecting this disease. Henceforth most of the previous research articles concentrated on detection of heart disease in at an early stage, as part of their medical pre-diagnosis research target.

To aid these medical diagnostic research operations, some social medical organizations made available their private medical data sets online, by changing the actual names and other personal information to protect the patient's data privacy. University of California, Irvine (UCI) online learning dataset repository provides the "Cleveland Heart Disease dataset"[6], which are donated by Dr. Robert Detrano[7]. This data set becomes a standard dataset due to its prominent medical properties and most of the cardiac researches are widely using this as a base.

Inspiring from the advanced machine learning technologies, former researches have employed different machine learning techniques like Decision Trees (DT), Support Vector Machines (SVM), Naïve Bayes (NB), Fuzzy Logics and K-Nearest Neighbor (KNN) etc. Each technique works in its unique way to classify the data and to extract the disease predictions. Ali Adil and M. Neshat [5] stated that current data mining techniques are suffering from imprecision, uncertainty and complexity in processing. High precision values in decision making returns the reliable results from medical diagnostic systems.

Motivated by the need of precision and reliability, in this paper, we are proposing a cross reference methodology to improvise the reliability and precision of diagnostic results and utilizing big data tools to diminish the complexity in processing huge sets of medical data. Present data mining techniques, which are participating in medical data processing with high accuracy are selected and cross referenced by our proposed framework to overcome uncertainty and imprecision. In order to process the high velocity medical datasets with several data mining techniques in parallel, this frame work outsources the data processing to Apache Hadoop environment. Experimental results from simulations represent that, our proposed framework with cross reference methodology helps in returning the high precision results for medical diagnostic systems. In order to accomplish the huge data set mining in less time, my framework had taken the help from Apache Hadoop with Map Reduce java programs.

Since few years, number of machine learning techniques has been used in the field of medical diagnosis systems development. In the year 1999, Azuaje et al [8] utilized Artificial Neural Networks (ANN) to recognize the variations in heart rate using pattern to determine the cardio vascular problems. Allahverdi.N et al[10]designed a Fuzzy Expert System in 2001, to pre-diagnosis the cardio vascular disease occurrence chances in future. Fuzzy logic expressions play a prominent action in fuzzy expert system to improvise the accuracy level of results. M. Gudadhe et al[9] proposed SVM based approach on Cleveland dataset[6] to diagnose the heart diseases based on data set contained medical records.

In this section, we discuss about the machine learning techniques, which are widely used in medical diagnosis like Decision Trees (DT), Support Vector Machines (SVM), Naïve Bayes (NB), Fuzzy Logics and K-Nearest Neighbor (KNN).

2.1 Decision Trees

Decision trees are mainly used in machine learning for classification and regression operations on given input data set. For decision making, initially data will be represented in the form of trees with root, descendants, ancestors, and siblings like relations. As decision tree is good in processing both numerical and character data, it can understand the medical data about a patient very easily. With the help of in-built variable screening and feature extraction capability, decision tree reduces the burden on client in providing formatted input data. As decision tree is flexible enough to apply on different domains without data preparation, we can use this on medical data set directly without any preprocessing burden of datasets.

2.2 Support Vector Machine (SVM)

SVM classifier is the most popular data classifier in machine learning. It works using the supervised learning model on labeled datasets by preparing the linear models while processing real time data. With the help of associated learning algorithms, SVM performs the classification and regression operations on big data sets. This is very impressive while using in classification of multi-dimensional data properties. As medical dataset is having total 13 different attributes about patient's health, which effects on disease diagnosis, SVM is a suitable classifier in design of pre-diagnostic systems. Using the training subsets, it can make the decisions with the help of hyper planes to reduce the burden of processing the data especially while dealing with larger datasets.

2.3 Naïve Bayes

This classifier is a group of data classification algorithms, which follows the common principle while dealing with dataset properties i.e., each node pair should be classified in an independent manner with each other. The given tabular data with labels divided into two sets are feature matrix with dependent features of data and response vector with decision making knowledge. Probability based Bayes Theorem helps in predicting the future analysis using naïve assumptions.

2.4 Fuzzy Logic

Fuzzy set theory based fuzzy logic is widely using in machine learning dependent applications like medicine, mechanical etc. Fuzzy logic creates the propositions and relations among data variables initially. It is having two different faces while dealing with medical datasets which are wide and narrow logics. Approximate inference design is possible at early stages of data classification using the reasoning methods of fuzzy logic. Using set theory elements $\wedge$ (and), $\vee$ (or), $'$ (not), fuzzy logic constructs the decision making mechanism.

2.5 K Nearest Neighbor (KNN)

By finding the nearest neighbor, among the group of neighbors with most relevant features, KNN returns more accurate results in a simplified manner. Distance calculation based on various properties with modern KNN algorithms is a small time consuming task. By adjusting

Copyright ©2019 GV Press

the threshold value we can expect the most accurate performance from this. Data will be trained using supervised training data and classified using the value of Euclidean distance. Hamming distance is used while dealing with categorical data to improve the results accuracy.

3. Accurate and Reliable Heart Disease Pre-Diagnosis Framework

In this section we represent the proposed Accurate and Reliable heart disease pre-diagnosis framework with respective components to overcome the imprecision, uncertainty and complexity in processing huge medical data sets. This framework consists of three interdependent layers which are Formatted Input data layer, cross reference validation model layer and fine grained result output layer. Apart from this another independent tier also part of this framework is Big data processing Hadoop environment.

3.1 Formatted Input Data Layer

This is the primary layer of designed framework, which is responsible to format the raw input data of medical datasets to machine learning algorithm process-able convenient data. In this process of data formatting, trimming, stemming and labeling operations are performed using popular data tokenization tools. Apart from data format changing, data arrangement structures also updated to make the medical dataset processing feasible. After this data preprocessing operation, training datasets are created with human intervention or existing training sets are insisted to support supervised learning techniques of data mining. Finally an MDSS system is used to transform the categorical patient information (with character and numerical data values) to mathematical variables. These mathematical variables are most convenient data values for manipulations and decision making by machine learning techniques. To improvise the reliability levels and to reach the research goal (reduce imprecision) we are partitioning the input data as patient's basic information, disease symptoms, medical history, lab reports, medication details, heredity disease info etc.

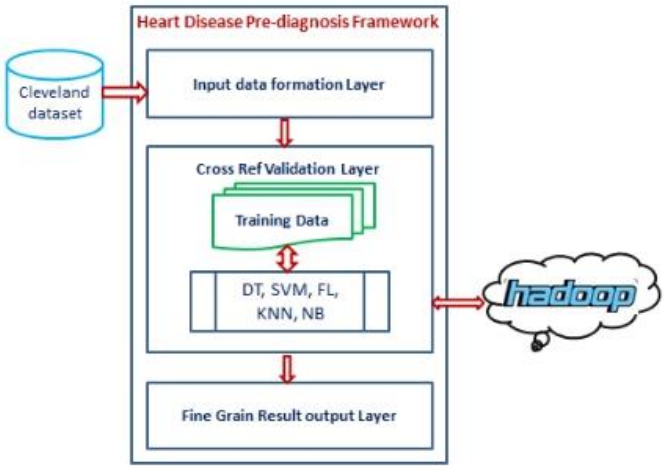

Figure 1: Architecture diagram of Accurate and Reliable heart disease pre-diagnosis framework

3.2 Cross reference validation layer

This is the second layer of proposed framework, which mainly concentrates on processing the data using several data mining techniques in parallel as shown in fig 1. Formatted Input

Data from input data layer will be loaded to the auxiliary storage system of this second layer for speed access of data like cache memory. This auxiliary data store connects with machine learning techniques layer through data transformation tool (DTT). DTT just tunes the coarse grained input data to Machine learning techniques like DM, SVM, and K-NN expected input model.

As specified in framework diagram figure 1, second layer of this framework is having a cluster of prominent machine learning techniques, which were widely used and succeeded in earlier research operations[5][8][9]. Behalf of finding the best machine learning technique among the available popular methods, in this paper to achieve the accuracy and reliability, we have strongly recommended, implemented and proved that, the parallel applicability of popular N methods on same medical data helps in breeding the best results. Because a single method cannot make a perfect decision as its view angle is limited to some scope or degree of consideration. As medical diagnosis systems are more concentrating on the accuracy and reliability of system results than its count, our research introduced this parallel execution of popular N machine learning methods on medical datasets to cover all degrees and to get more accurate results with confidence.

As our framework performs cross reference operation against popular machine learning methods using a common dataset, handling the processor and memory needs become too complex due to heavy loading using traditional mainframes computers with static configuration. To overcome the problem of load balancing and to alleviate the burden of infrastructure management, we adopted the cloud based big data processing environment [11] as part of cloud service PaaS. This environment was implemented on Google Cloud Platform [12] with Apache Hadoop framework [13]to balance the load efficiently and to perform the complex medical data analysis with learning methods in very less time. By using distributed file systems and map reducing techniques of big data, we implemented our analysis algorithm of each learning method in java. This environment helps in load balancing and speed processing of huge medical datasets without infrastructure management burdens. Process execution flow and the sample dataset results are explained in detail in experimental analysis section of this paper. After processing the dataset using popular N learning methods, results returned by each method will be forwarded to the fine grained result output layer.

3.3 Fine grained result output layer

This is the bottom layer of the proposed framework, which is responsible to return the fine grained results (pre diagnosis disease information) to target audience like doctors, medical practitioners, medical research scholars etc. Once the results of each learning method on common medical data set received by this layer from cross reference validation model layer, it performs the commonality and probability operations on each method recommendations to evaluate the reliable and accurate diagnosis information. The recommendations with more commonality considered as “complete reliable information”, at the same time recommendations with more probability also returned as “maximum reliable information”. Medical staff uses this information and starts the activities to prevent the heart disease at first hand. The maximum possible probability based diagnosed diseases are monitored periodically by medical staff with physical examinations (lab reports) to prevent them. In this way the proposed framework, not only is best in diagnosing the assured occurrence of diseases, but also helps in keeping an eye on future possibilities.

6. Conclusion and Future Work

Copyright ©2019 GV Press

To overcome the imprecision, uncertainty and data processing problems of today's medical diagnosis systems, we proposed a cross reference methodology to improvise the reliability and precision of diagnostic results, and introduce to utilize the big data tools to diminish the complexity in processing huge sets of medical data. To implement the proposed methodology we designed Accurate and Reliable heart disease pre-diagnosis framework with the support of cloud based Hadoop platform for mining huge medical dataset to extract pre-diagnostic results with accuracy. Experiments on Cleveland dataset proved that the proposed cross reference methodology framework recorded high precision, recall and accuracy, which denotes the accuracy and reliability of results. In future, we are planning to extend this research with different goals like finding the best machine learning classifiers combination for medical data processing with high accuracy, and implementing data and resource reusability among machine learning classifiers to minimize the redundant processing of data.

References

[1] S. Palaniappan and R. Awang, “Intelligent heart disease prediction system using data mining techniques,” in IEEE/ACS International Conference on Computer Systems and Applications. IEEE, **(2008)**, pp.108-115, [DOI:10.1109/AICCSA.2008.4493524].

[2] Azuaje, F., Dubitzky, W., Lopes, P., Black, N., & Adamsom, K. “Predicting coronary disease risk based on short-term RR interval measurements: A neural network approach”. Artificial Intelligence in Medicine, Vol.15, No.3, pp.275-297, **(1999)**. [DOI:10.1371/journal.pone.0210103]

[3] Randa El Bialy, Mostafa, Omar and Khalifa “Feature analysis of coronary heart disease data sets” ICCMIT – 2015, Vol.65, by Elsiever’s science direct, procedia comp science, pp.459-468. **(2015)** [DOI: 10.1016/j.procs.2015.09.132]

[4] Indira S. Fal Dessai “Intelligent Heart Disease Prediction System Using Probabilistic Neural Network”, International Journal on Advanced Computer Theory and Engineering (IJACTE), ISSN (Print) : pp.2319-2526, Vol.2, No.3, **(2013)**

[5] Ali.Adeli, Mehdi.Neshat “A Fuzzy Expert System for Heart Disease Diagnosis” Proceedings of the international multi conference of engineers and computer scientists, vol. 1, March-2010, hongkong.**(2010)**

[6] University of California, Irvine (UCI) “Online accessible public medical data set: Cleveland Heart Disease Dataset,”. [Online] Available at : http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[7] Robert Detrano & M.D & PhD, V.A. Medical Center, Long Each and Cleveland Clinic Foundation. Available: www.archive.ics.uci.edu/ml/datasets/Heart+Disease.

[8] Azuaje, F., Dubitzky, W., Lopes, P., Black, N., & Adamsom, K. “Predicting coronary disease risk based on short-term RR interval measurements: A neural network approach”. Artificial Intelligence in Medicine, 15, pp.275-297, **(1999)**. [DOI:10.1023/B:JOMS.0000041169.28544.fd].

[9] M. Gudadhe, K. Wankhade, and S. Dongre, “Decision support system for heart disease based on support vector machine and artificial neural network,” in Computer and Communication Technology (ICCCT), 2010 International Conference on, **(2010)**, pp. 741-745, [DOI: 10.1109/ISCC.2017.8024530].

[10] H. Kahramanli and N. Allahverdi, “Design of a hybrid system for the diabetes and heart diseases,” Expert systems with applications, vol. 35, no. 1, pp. 82-89, **(2008)**, DOI: 10.1016/j.eswa.2007.06.004.

[11] Samiya Khan, Kashish Ara Shakil and Mansaf Alam “Cloud-Based Big Data Analytics – A Survey Of Current Research And Future Directions” Big Data Analytics. Advances in Intelligent Systems and Computing, vol.654. Springer, Singapore,**(2017)** [DOI: 10.1007/978-981-10-6620-7_57]

[12] Google Cloud Platform and Big Query. Retrieved from: https://cloud.google.com/bigquery/

[13] K. Shvachko, H. Kuang, S. Radia and R. Chansler, "The Hadoop Distributed File System," 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), Incline Village, NV, **(2010)**, pp.1-10. [doi: 10.1109/MSST.2010.5496972]