Human Healthcare System using KNN and Genetic Algorithm

Sayali Avhad [#1], Rupali Borate[#2], Shweta Borate [#3], Charushila Gaikwad [#4], Shraddha Toney (Guide) [*5]

[#]Department of Computer Engineering, Sinhgad Institute of Technology & Science, Narhe, Pune

[1]sayaliavhad1995@gmail.com

[2]borateruplali5@gmail.com

[3]shwetaborate1995@gmail.com

[4]charushilagaikwad33@gmail.com

[*] *Department of Computer Engineering, Sinhgad Institute of Technology & Science, Narhe, Pune*

[5]*toneyshraddha@gmail.com*

***Abstract*— Future HealthCare must be hold for people around the world. The goal of the system is to prevent people to go to hospital, in other word prevent to be sick. HealthCare system will do a pre-check, and follow up with how to take care of your body, an example weather issues can prevent someone to go somewhere. Technology within the system is how can we bring all the symptoms checker to our closely device and give the summary of your present body. Data mining techniques have been widely used to mine knowledgeable information from the medical databases. In data mining Classification is supervised learning that can be used to design models describing important data classes, where class attribute is involved in construction of classifier. Nearest neighbours is very simple, most popular, highly efficient and effective algorithm for pattern recognition. KNN is a straight forward classifier, where sample are classified based on the class of their nearest neighbour. Medical databases are high volume in nature. If the data set contains redundant and irrelevant attributes, classification may produce poor result.**

Key words: KNN, GA, SVM

I. Introduction

Data mining is the process of extracting knowledgeable information from huge amount of data. It has become increasingly important as real life data enormously increasing. The basic functionality of data mining involves Classification, Clustering and Association. Classification is a pervasive problem that encloses many diverse applications. To improve medical decision making techniques have been applied to variety of medical domains. Data mining techniques answered several important and critical questions related to health care. Nearest neighbor is one of the most popular classification technique. Without any additional information, classification rules are generated by the training data samples themselves. K- Nearest neighbor (KNN) is a simple algorithm, which stores all the cases and classify new cases based on similarity measure. KNN algorithm has been used since 1970 in many applications like pattern recognition and statistical estimation. Genetic algorithms are most popular technique in evolutionary computing. Evolutionary algorithms are used in problems for optimization. To solve problems, evolutionary algorithms require a data structure to represent and evaluate optimized solution from set of solutions. Along with the KNN algorithms GA has been applied to find an optimal set of feature weights that improve classification accuracy.

II. Literature survey

According to M. Akhil jabbar et. all [1] Heart disease is one of the main contributor to disease burden in developing countries like INDIA and the leading cause of death in developed countries. Hence there is a need to design and develop a clinical decision support for classification of heart disease. In this paper KNN and genetic algorithm are combined to predict heart disease of a patient for Andhra Pradesh population. In this paper, proposed approach combines KNN and genetic algorithm in order to improve the classification accuracy of heart disease data set. They used genetic algorithm which ignores redundant and irrelevant attributes and rank the attributes which contributes more towards classification. This classifier is trained to classify disease data set as either healthy or sick. Proposed method is tested against 6 medical datasets and 1 non-medical dataset which were chosen from UCI repository. Heart disease A.P was taken from various corporate hospitals in Andhra Pradesh, and attributes are selected based on opinion from expert doctor's advice, then 5 fold cross validation is applied on the dataset. A new algorithm which combines KNN with genetic algorithm gives effective classification for Heart disease data set. Proposed method (KNN+GA) was not successful for breast cancer and primary tumor. This prediction model helps the doctors in efficient heart disease diagnosis process with less attributes.

According to Rusdah et. all [2] Tuberculosis is one of oldest human disease which is mainly knows as an infectious disease. Tuberculosis is a bacterial infection which causes more diseases in the world than any other infectious diseases.

Diagnosis of tuberculosis is difficult because the tuberculosis symptoms which are not similar to lung cancer but also with the other type of diseases. Hence detection of Tuberculosis is difficult. Preliminary diagnosis can be established by using patient demographic data, anamnesis and physical examination. And also some experiments are conducted using classification technique also which includes C4.5, Naïve Bayes, Backpropagation and SVM in order to improve the performance of the model. Research method proposed in the paper in

- Data source

Data used in research and diagnosis were patient's real data taken from JRC (Jakarta Respiratory Centre) from 2010 to 2014. The data consists of 1170 records having 17 attributes. Those attributes are age, sex, village, sub district, occupation, duration of cough, fever, weight loss, chest pain etc.

Name of Attribute	**Data types**
Age	N
Sex	C
Village	C
Occupation	C
Duration of cough	C
Fever	C
Chest pain	C
Weight	N

- Data exploration

According to their research there are 535 cases of tuberculosis, 587 cases of non-TB and 48 case of extra pulmonary tuberculosis. Males were 56.8% of total cases which is higher as compared to females. Among the data used for diagnosis, there are 0.5% cases of missing values**.**

- Data preprocessing

In Study they have used data preprocessing techniques like data cleaning for handling missing values, data transformation and data discretization of numerical values

- Modeling

For first experiment we used RM2_17 datasets and classify the dataset using four classifiers, C4.5, Naïve Bayes, Back-propagation, and SVM. We trained the dataset using 10 fold cross-validation. The results showed that SVM had the highest accuracy of 66.92%, followed by Naive Bayes, Back-propagation and C4.5. The datasets trained using 10 fold cross validation and classified using ensemble methods, namely Bagging and Boosting. The methods combined other single classifiers, i.e. C4.5, Naive Bayes, Back-propagation, and SVM. Bagging is found to be better than Boosting in both of datasets. Ensemble method improve classification accuracy. Ensemble methods performs better than single classifier.

According to Mr. Jaykurnar Lachure et. all [3]. Diabetic Retinopathy that is DR which is a eye disease that affects retina and further it leads to vision loss. Early detection of DR is helpful to improve the screening of patient to prevent further damage. The main aim of proposed work is to detect retinal micro aneurysms and exudates for automatic screening of DR using Support Vector Machine (SVM) and KNN classifier. To develop this proposed system, a detection of bright and red lesions in digital fundus photographs is needed. In given proposed methodology as shown in Fig. 1 shows that image is get input from given data set pre-processing methods are applied then morphological operations are performed to identify exudates and micro-aneurysms. Finally, by applying multi class SVM and KNN classifier giving severity or degree of abnormality. For given methodology the input images are taken from MESSIDOR, Diabetic ret DB 1. The input image of size 2240 X 1488 pixels in .tiff format is used for pre-processing stage. In preprocessing stage, the image is get rectified from the problems such as blurring, non-clarity and size. After preprocessing detection of Micro-aneurysms and Exudates takes place. After detecting exudates and micro-aneurysms in color fundus image, the features get extracted. Splat, GLCM (Gray Level Co-occurrence Matrix) and calculated are applied to SVM and KNN classifier. SVM classifier gives better results than KNN classifier. So from the extracted feature it directly concludes the disease grad as normal, moderate and severe. So earlier detection and diagnosis of Diabetic retinopathy help the patients from vision loss and also the severity of disease can be decreases.

According to Veenita Kunwar et. all [4]. The present lifestyle of people, working environment and diet have given rise to many diseases, one of which includes chronic kidney disease. Chronic Kidney disease (CKD) is prevailing nowadays and has become a global health issue which must be timely detected and diagnosed. CKD is a condition that describes loss of kidney function over time making it difficult for them to filter poisonous wastes from the body. It has been observed that classification algorithms have widely been used for identifying and investigating kidney disease. Chronic Kidney Disease has been predicted and diagnosed using data mining classifiers: ANN and Naive Bayes. Performances of these algorithms are compared using Rapid miner tool.

Data Set:

The clinical data of 400 records considered for analysis has been taken from UCI Machine Learning Repository. The data obtained after cleaning and removing missing values is 220. The data has been implemented using Rapid Miner tool. There are 25 attributes in the dataset. The numerical attributes include age, blood pressure, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packaged cell volume, WBC count, RBC count. The nominal attributes include specific gravity, albumin, sugar, RBC, pus cell, pus cell clumps, bacteria, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anemia and class. The experimental comparison of Naive bayes and ANN

are done based on the performance vectors. It is statistical performance evaluation of classification tasks and contains list of performance criteria values. The obtained results showed that Naive Bayes is the most accurate classifier with 100% accuracy when compared to ANN having 72.73% accuracy. In this research study, some of the factors considered were age, diabetes, blood pressure, RBC count etc. The work can be extended by considering other parameters like food type, working environment, living conditions, availability of clean water, environmental factors etc. for kidney disease detection.

III. Proposed Methodology

- Basic Idea

The proposed classification algorithm combines KNN and genetic search algorithm, to predict diseases of a patient, in which genetic search is applied on training data set to check redundant and irrelevant attributes and to rank the attributes which contribute towards the better classification. Lower ranked attributes are neglected , and classification algorithm is built based on evaluated attributes. The classifier classifies the disease data set as sick.

- Algorithm

1. Apply genetic search on the data set.
2. Attributes are ranked based on their value.
3. Select the subset of higher ranked attributes.
4. Apply (KNN+GA) on the subset of attributes that maximizes classification accuracy.
5. Calculate accuracy of the classifier.

The idea in k-Nearest Neighbour algorithm is to identify k attributes in the training data set that are similar to a new attribute , say (u1, u2, ...up), that we wish to classify and to use these observations to classify the observation into a class, v.If we knew the function f, we would simply compute v = f (u1, u2,…..up).

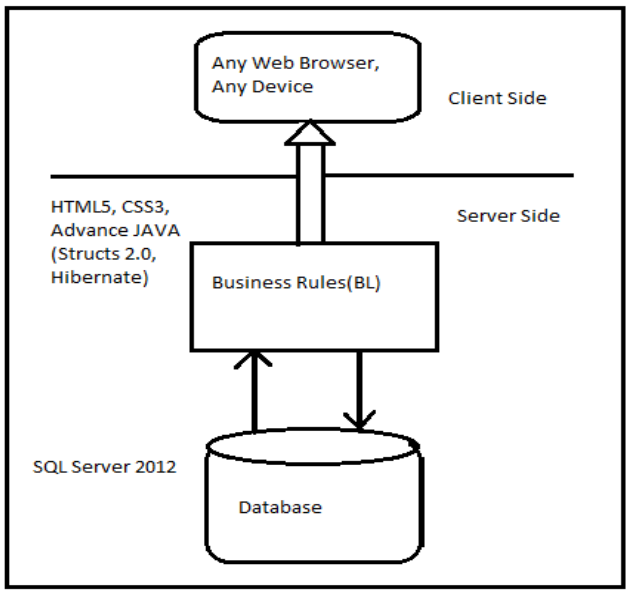

Fig1 System Architecture

Fig1 shows the basic system architecture of proposed idea. It follows three tier architecture. Front end consists of actual device with a web browser. At the back end we have Business logic (BL) and Database. Database contains primary information of users, Disease Symptoms and relevant diseases. We used SQL Server 2012 for storing the data. Using Java as programming language we are going to implement business logic of proposed system.

IV. Conclusion

This application can be easily implemented under different situations. In current fast growing world, time is important for each & every person. So, we are trying to reduce wasting time of patient to wait in queue at hospital. This prediction model helps the doctors in efficient Human disease diagnosis process with fewer attributes. We can add new features as per user's requirement. . As the accuracy resulted in all experiments is not too high, a future work to improve the accuracy is still needed. We have done survey on different data mining algorithms and techniques for disease diagnosis, still No method is 100% efficient for disease diagnosis.

References

[1] [1] M.Akhil jabbar, B.L Deekshatulu, Priti Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm", International Conference on Computational Intelligence: Modeling Techniques and Applications (C IMTA) 2013.

[2] [2] Rusdah, Edi Winarko, Retantyo Wardoyo, "Preliminary Diagnosis of Pulmonary Tuberculosis Using Ensemble Method", International Conference on Data and Software Engineering 2015.

[3] [3] Mr. JaykurnarLachure , Mr. A. V .Deorankar, Mr. Sagar Lachure, Miss. Swati Gupta, Mr. Romit Jadhav "Diabetic Retinopathy using Morphological Operations and Machine Learning" IEEE 2015.

[4] [4] Veenita Kunwar, Khushboo Chandel, Sai Sabitha, Abhay Bansal4, "CHRONIC KIDNEY DISEASE ANALYSIS USING DATA MINING CLASSIFICATION TECHNIQUES" IEEE 2016.

[5] [5] G.Sumalatha, Dr.N.J.R. Muniraj, "Survey on Medical Diagnosis Using Data Mining Techniques", IEEE 2013.

[6] [6] Agarwal, Y., & Pandey, H. M., "Performance evaluation of different techniques in the context of data mining-A case of an eye disease. In Confluence the Next Generation Information Technology Summit (Confluence)", IEEE 2014.

[7] [7] Aanchal Oswal, Vachana Shetty, Mustafa Badshah, Rohit Pitre, Manali Vashi, "A SURVEY ON DISEASE DIAGONSIS ALGORITHMS", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 11, November 2014.

[8] [8] Nitin Bhatia, vandana, "Survey on nearest neighbor techniques" IJCSIS, Vol 80, no 2(2010)

[9] [9] MA.Jabbar, B.L Deekshatulu, Priti chandra,"Heart disease prediction system using associative classification and genetic algorithm" pp183-192 Elsevier (2012)

[10] [10] MA.Jabbar, B.L Deekshatulu, Priti Chandra,"An evolutionary algorithm for heart disease prediction" CCIS, PP 378-389, Springer(2012)

[11] [11] MA.Jabbar, B.L Deekshatulu, Priti chandra, "Prediction of Risk Score for Heart Disease using Associative classification and Hybrid Feature Subset Selection", IEEE(2013).

[12] [12] T. Uçar, A. Karahoca, dan D. Karahoca, "Tuberculosis disease diagnosis by using adaptive neuro fuzzy inference system and rough sets," Neural Computing and Applications, vol. 23, no. 2, pp. 471–483, Apr. 2013.

[13] [13] T. Asha, S. Natarajan, dan K. N. B. Murthy, "Data Mining Techniques in the Diagnosis of Tuberculosis," in Understanding Tuberculosis - Global Experiences and Innovative Approaches to the Diagnosis, P.-J. Cardona, Ed. InTech, 2012, pp. 333 – 352.

[14] [14] F. S. Aguiar, L. L. Almeida, A. Ruffino-Netto, A. L. Kritski, F. C. D. Q. Mello, dan G. L. Werneck, "Classification and regression tree (CART) model to predict pulmonary tuberculosis in hospitalized patients.," BMC pulmonary medicine, vol. 12, no. 1, pp. 12 – 40, Jan. 2012.