Fuzzy & Datamining based Disease Prediction Using K-NN Algorithm

Anand A. Chaudhari
Computer Science & Engineering Department
PRMIT&R, Badnera-Amravati, Maharashtra, INDIA

Prof.S.P.Akarte
Computer Science & Engineering Department
PRMIT&R, Badnera-Amravati, Maharashtra, INDIA

Abstract- Disease diagnosis is one of the most important applications of such system as it is one of the leading causes of deaths all over the world. Almost all system predicting disease use inputs from complex tests conducted in labs and none of the system predicts disease based on the risk factors such as tobacco smoking, alcohol intake, age, family history, diabetes, hypertension, high cholesterol, physical inactivity, obesity. Researchers have been using several data mining techniques to help health care professionals in the diagnosis of heart disease. K-Nearest-Neighbour (KNN) is one of the successful data mining techniques used in classification problems. However, it is less used in the diagnosis of heart disease patients. Recently, researchers are showing that combining different classifiers through voting is outperforming other single classifiers. This paper investigates applying KNN to help healthcare professionals in the diagnosis of disease specially heart disease. It also investigates if integrating voting with KNN can enhance its accuracy in the diagnosis of heart disease patients. The results show that applying KNN could achieve higher accuracy than neural network ensemble in the diagnosis of heart disease patients. The results also show that applying voting could not enhance the KNN accuracy in the diagnosis of heart disease.

Keywords – Data Mining, Neural Network, K-NN, Genetic Algorithm

I. INTRODUCTION

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Jyoti Soni et.al [3] proposed three different supervised machine learning algorithms. They are Naïve Bayes, K-NN, and Decision List algorithm. These algorithms have been used for analyzing dataset [14].

Decision tree is one of the popular and important classifier which is easy and simple to implement. It doesn't have domain knowledge or parameter setting. It handle huge amount of dimensional data. It is more suitable for exploratory knowledge discovery. The results attained from Decision Tree are easier to interpret and read [7]. Naïve Bayes is a statistical classifier which assigns no dependency between attributes. To determine the class the posterior probability should be maximized. The advantages are one can work with the naïve bayes model without using any Bayesian methods. Here Naïve Bayes Classifiers performs well [1].

Motivated by the world-wide increasing mortality of heart disease patients each year and the availability of huge amount of patients' data that could be used to extract useful knowledge, researchers have been using data mining techniques to help health care professionals in the diagnosis of heart disease [5-6]. Data mining is an essential step in knowledge discovery. It is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to be detected with traditional statistical methods [7-11]. The application of data mining is rapidly spreading in a wide range of sectors such as analysis of organic compounds, financial forecasting, healthcare and weather forecasting [12].

Data mining in healthcare is an emerging field of high importance for providing prognosis and a deeper understanding of medical data. Healthcare data mining attempts to solve real world health problems in diagnosis and treatment of diseases [13]. Researchers are using data mining techniques in the medical diagnosis of several diseases such as diabetes [14], stroke [15], cancer [16], and heart disease [17]. Several data mining techniques are used in the diagnosis of heart disease showing different levels of accuracy. K-Nearest-Neighbour (KNN) is one of the most widely used data mining techniques in pattern recognition and classification problems [18]. Recently Paris et al. examined single classifiers and combining different classifiers through voting and showed that voting outperformed other single classifiers [19]. This paper investigates applying KNN in the diagnosis of heart disease on the benchmark dataset to allow comparisons with other data mining techniques used on the same dataset. It also investigates if integrating

II. Implementation Of Algorithm

A. Motivation K-NN Algorithm

Jyoti Soni proposed three different supervised machine learning algorithms [7]. These are Naïve Bayes, K-NN, and Decision List algorithm. These algorithms have been used for analyzing the heart disease dataset. Decision tree is one of the popular and important classifier which is easy and simple to implement. It does not have domain knowledge or parameter setting. It handle huge amount of dimensional data. It is more suitable for exploratory knowledge discovery. The results attained from Decision Tree are easier to interpret and read. Naïve Bayes is a statistical classifier which assigns no dependency between attributes. K-nearest neighbor's algorithm (k-NN) is the one of the important method for classifying objects based on closest training data in the feature space. It is simplest among all machines learning algorithm but, the accuracy of k-NN algorithm can be degraded by presence of noisy features. The dataset is divided into two testing and training i.e. 70% of data are used for training and 30 % is used for testing. The authors concluded that Naïve Bayes algorithm performs well when compared to other algorithms. In the survey of Naïve bayes have been used to predict attributes such as age, sex, blood pressure etc.

B. K-NN Algorithm

This method of classification is one of the most fundamental and simple classification methods and should be used for a classification study when there is little or no prior knowledge about the distribution of the data. This method was developed from the need to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine [8].
The algorithm for this method is

- The k nearest neighbor must be located using the training dataset. The Euclidean distance measure is used to calculate how close each member of the training set is to the target row that is being examined.
- Examine the k- nearest neighbor, which classification or category do most of them belong to? Assign this classification or category to the row being examined.
- Repeat this procedure for the remaining rows in the target set.
- In this software a maximum value for k can be selected, then the software builds models parallel on all values of k up to the maximum specified value and scoring is done on the best of these models.

The First step in using K-Nearest Neighbor classification method in weka software was determining the training data set, then the input and output variables should be entered. The second step was normalizing the data which will ensure that the distance measure accords equal weight to each variable. The score on best k between 1 and specified value was chosen which builds models parallel on all values of k up to the maximum specified value in which k=9 was chosen and scoring is done on the best of these models. Finally entering the data needed for classification.

C. Fuzzy Based Approach

Now, let us deal with "fuzzy logic" in medicine in broad sense. In the medicine, especially, in oriental medicine, most medical concepts are fuzzy. The imprecise nature of medical concepts and their relationships requires the use of "fuzzy logic". It defines inexact medical entities as fuzzy sets and provides a linguistic approach with an excellent approximation to texts. "Fuzzy logic" offer reasoning methods capable of drawing approximate inferences.

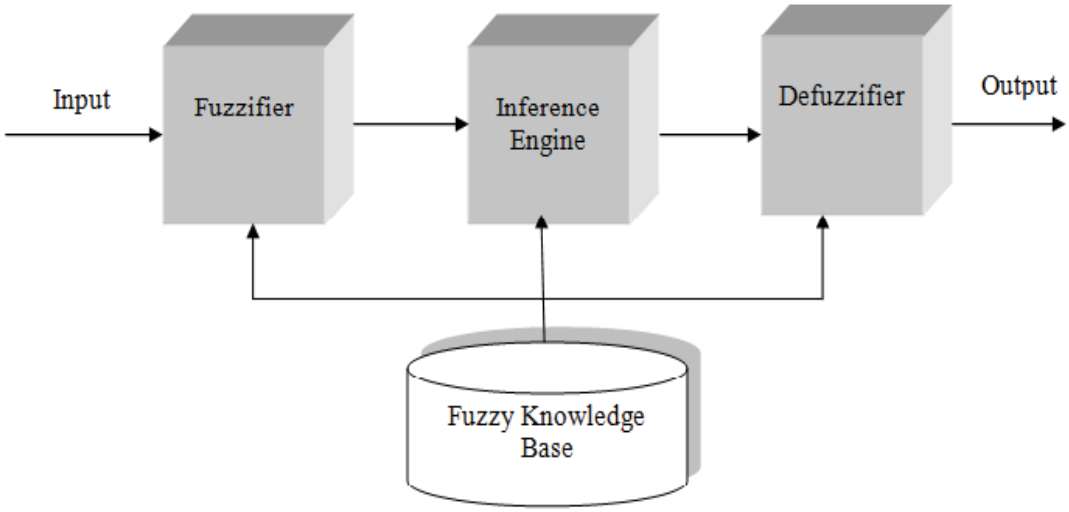

Figure 1: Fuzzy Logic extraction

D. Rule Based fuzzy systems in medicine

In rule-Based fuzzy systems in medicine, experts often formulate their statement in terms of rules of the type:

If x is A and y is B then z is C

For example,
If back pain is severe and patient is old then apply acupuncture to a certain point for a long time.

Here:
X is patient's pain, A is "severe";
y is patient's age, B is "old"
z describe treatment's time, C is "long time"

Now, we describe the formalization of such if-then rules in the rule base of expert systems. For each rule:
If x1 is A1, ..., xn is An then z is C
We can compute the degree to which the conditions are applicable as
μcond = μ A1(x1) ^...^ μ An(xn)

Them, for each possible z, we can compute the degree to which the rule holds:
μrule = μcond.^μ C(z).

III. Experiment and Result

Methodology I:
To detect diabetes - A classifier is a detection function which classifies between two or more classes by assigning labels to the individuals.
Specific to our project, classifier can be defined as -
F: D8 → R

Where,
1. F is the function that matches the domain inputs to the range outputs.
2. D8 is the 8 dimensional input-domains that has following the 8 attributes that are present in Pima Indian Database
   i. No. of times pregnant
   ii. Plasma Glucose concentration a 2 hrs oral glucose tolerance test

iii. Diastolic Blood Pressure (mmHg)
iv. Triceps Skin fold thickness (mm)
v. Hour serum insulin (mu U/ml)
vi. Body Mass Index
vii. Diabetes Pedigree Function
viii. Age

3. R is the range i.e. {Diabetic, Non-diabetic}

The proposed system will use a Neural Networks for design of a classifier for detection of diabetes. Neural Networks are popular for their dynamic nature in terms of learning, which is why they are preferably chosen for medical diagnosis. In spite of having such a great quality, the neural networks cannot predict with a remarkable accuracy. The fuzzy systems, though not as dynamic as neural networks, can work accurately owing the fact that they have rule bases that mimic that human thinking. When neural networks are made to control the rules generated by the fuzzy systems, the neuro-fuzzy systems are created. Instead of modifying the fuzzy systems as per the predictions of the neural networks, the presented approach focuses on implementing the neural networks as a fuzzy system.

This method of detection of diabetes proposes a system that will be implemented in client-server architecture. Here, the training dataset will be kept on the server, which will be used to train the neural network classifier on the mobile device. The mobile device is a feature add-on for convenience of the doctor. The relevant processing of the input data will be done on the server to spare the adverse effect on the performance of the device. The processing will include the search for k nearest neighbors using k-NN algorithm and fuzzy allotment of class for the input. The neural networks will be made to do this implementation. The accuracy of this proposed method is calculated to be 72.8281% for 10 fold CV in WEKA classifier. The accuracy surges to 100% when all the attributes are known and the training set is used as the test set. Owing to erroneous dataset of the UCI Pima Indians Diabetic Dataset, removal of records with missing values is considered. With removal of various attributes that are less significant in the context, the accuracy of classification ranges between 75% and 100%. First of all, the proposed model eliminates the records containing the missing values from the Pima Indian Diabetes Dataset. Based from the literature survey [10], it is dimension of the dataset are reduced to half of the previous. The Fuzzy k- Nearest Neighbor algorithm is used to train the Neural Networks. Finally, the entire training set is used as test set to calculate the classification accuracy.

ID	SEX	Diastolic B.P mm Hg	Plasma glucose mg/dL	Skin fold thick mm	BMI Kg/m^2	Diabetes Pedigree type	No. of times pregnant	2 hr Serum Insulin muU/ml	Diabetes probability
1	F	100	182.5	27.76	31.75	2	0	140	High
2	M	68	98.30	35.75	28.12	1	-	54	Low
3	M	88	111.36	35.25	28.95	2	-	78	Low
4	F	52	131.18	27.68	28.75	2	1	122	Medium
5	F	73	142.2	28.64	28.55	1	0	105	Medium

Table -1 Experiment Result for Diabetes Prediction

Methodology II:

Heart Disease Risk Factor: Risk factor for a pataient can be calculated by passing various parameters such as weight, sex disease relate sysmtoms, BMI ratio, Food Intake Type etc. During the experiment it was found that KNN works better as compared to genetic algorithm for detecting the risk factor of heart disease for coming years.

Table -2 Experiment Result for Heart Disease Prediction Risk

Years	KNN	Genetic Algorithm
After 4 years	**7 %**	**3.5 %**
After 8 years	**12%**	**11 %**
After 10 years	**15 %**	**22%**

It was found that KNN gives more accurate prediction as compared to K-NN Algorithm.

IV.CONCLUSION

The accuracy of the proposed system is found to be good for both heart disease databases with KNN Algorithm when compared to that of the existing work. With the prediction of coronary heart disease, early treatment can be given at the right time which avoids the risk of heart attacks. Since the diagnosis involves simple procedures and is easy to obtain the required results, the proposed system is found to be efficient than the other existing systems.In this Paper the KNN algorithm of data mining is discussed for disease prediction. KNN algorithm proves to be more accurate when implemented with fuzzy rules not for all disease but for certain set of disease. Also we can say that KNN works even better when we have quality datasets available and the mode of input selected. The main focus is on using combination of data mining algorithm and combining several targets attributes for different types of disease prediction.

REFERENCES

[1] Kavita Agrawal, Syed Umar Amin, "Genetic Neural Network Based Data Mining In Prediction Of Heart Disease",2013.

[2] Farhad Soleimanian Gharehchopogh, "Neural Network Application In Diagnosis Of Patient", 2011.

[3] Sellappan Palaniappan, Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques", 2008.

[4] Ah Chen, Sy Huang, Ej Lin, "Heart Disease Prediction System",2011.

[5] M. Akhil Jabbar, Priti Chandra, "Prediction Of Risk Score For Heart Disease"2012.

[6] [Pang-Ning Tan,Michael Steinbach, "Introduction To Data Mining", 2008.

[7] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni, "Predictive Data Mining For Medical Diagnosis: An Overview Of Heart Disease Prediction" Ijcse Vol. 3 No. 6 June 2011.

[8] Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg, "Genetic Neural Network Based Data Mining In Prediction Of Heart Disease Using Risk Factors", 2013.

[9] Guo-Cheng Lan, Chao-Hui Lee, Yu-Yen Lee, Chu-Yu Chin, Miin-Luen Day, Shyh-Chyi Wang, "Disease Risk Prediction By Mining Personalized Health Trend Patterns: A Case Study On Diabetes", 2012.

[10] Minas A. Karaolis, Joseph A. Moutiris, Demetra Hadjipanayi, "Assessment Of The Risk Factors Of Coronry Heart Events Based On Data Mining With Decision Trees",2010.

[11] M. Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar, "Enhanced Prediction Of Heart Disease With Feature Subset Selection Using Genetic Algorithm", 2010.

[12] Asha Rajkumar, Mrs. G.Sophia Reena, "Diagnosis Of Heart Disease Using Data Mining Algorithm", 2010.

[13] K.Srinivas B.Kavihta Rani Dr. A.Govrdhan, "Applications Of Data Mining Techniques In Healthcare And Prediction Of Heart Attacks", 2010.

[14] Shantakumar B.Patil, Dr. Y. S. Kumaraswamy, "Extraction Of Significant Patterns From Heart Disease Warehouses For Heart Attack Prediction", 2009.

[15] Tahseen A. Jilani, Huda Yasin, Madiha Yasin, Cemal Ardil, "Acute Coronary Syndrome Prediction Using Data Mining Techniques- An Application", 2009.

[16] Asil Oztekin, Dursun Delen, Zhenyu (James) Kong, "Predicting The Graft Survival For Heart–Lung Transplantation

[17] Patients: An Integrated Data Mining Methodology", 2009.

[18] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", 2008.

[19] Latha Parthiban And R.Subramanian, "Intelligent Heart Disease Prediction System Using Canfis And Genetic Algorithm", 2008.

[20] Markos G. Tsipouras,Themis P. Exarchos, Dimitrios I. Fotiadis, Anna P. Kotsia, Konstantinos V. Vakalis, Katerina K. Naka, And Lampros K. Michalis, "Automated Diagnosis Of Coronary Artery Diseasebased On Data Mining And Fuzzy Modeling", 2008.

[21] Resul Das, Ibrahim Turkoglu, Abdulkadir Sengur, "Effective Diagnosis Of Heart Disease Through Neural Networks Ensembles", 2008.

[22] S.Kalaiarasi Anbananthen, G.Sainarayanan, Ali Chekima, Jason Teo, "Data Mining Using Artificial Neural Network Tree", 2005.

[23] Carlos Ordonez, "Comparing Association Rules And Decision Treesfor Disease Prediction", 2006.

[24] Jason H. Moore, Joshua C. Gilbert, Chia-Ti Tsai, Fu-Tien Chiang, Todd Holden, Nate Barney, Bill C. White, "A Flexible Computational Framework For Detecting, Characterizing, And Interpreting Statistical Patterns Of Epitasis In Genetic Studies Of Human Disease Susceptibility", 2006.

[25] Dragan Gamberger, Nada Lavrac, Goran Krstac, "Active Subgroup Mining: A Case Study In Coronary Heart Disease Risk Group Detection", 2003.

[26] U. Rajendra Acharya, P. Subbanna Bhat, S.S. Iyengar, Ashok Rao, Sumeet Dua, "Classification Of Heart Rate Data Using Artificial Neural Network And Fuzzy Equivalence Relation", 2003.

[27] Erkki Vartiainen, Juha Pekkanen, Seppo Koskinen, Pekka Jousilahti, Veikko Salomaa, Pekka Puska, "Do Changes In Cardiovascular Risk Factors Explain The Increasing Socioeconomic Difference Immortality From Ischemic Heart Disease In Finland?", 1998.