Investigation of the Diseases Prediction Rate using Specified Heuristic Method

[1]Manzar Haidar, [2]Prof. Sarvesh Site

M. Tech. Scholar, Department of Computer Science and Engineering, All Saints' College of Technology, Bhopal[1]

Guide, Department of Computer Science and Engineering, All Saints' College of Technology, Bhopal[2]

Abstract: The mining of health care data is important aspect for the prediction and estimation of critical disease of previous record. For the process of health care data mining various tools and technology are used. In concern of technology data mining algorithm are used. There has been increasing interest in gathering nontraditional, digital information to perform disease surveillance. These include diverse datasets such as those stemming from social media, internet search, and environmental data. With big data growth in biomedical and healthcare communities, accurate analysis of medical data benefits early disease detection, patient care, and community services. In this paper we proposed a new model which is based on the classification methods such as k nearest neighbor classification, decision tree classification with the optimization methods such as swarm intelligence family methods i.e. particle swarm optimization. Proposed optimization methods provide the better classification rate for features data using diseases prediction. Here we measure the accuracy for the various diseases data set such as heart dataset, Cleveland dataset and diabetes dataset, all the datasets are extracted from the uci machine learning repository and the simulation is done with MATLAB software.

Keywords: - Disease Prediction, Diabetes Dataset, Machine Learning, MATLAB Simulation

I. INTRODUCTION

The term “big data” has become a buzzword in recent years, with its usage frequency having doubled each year in the last few years according to common search engines. Below figure exemplifies the rapid rise in the number of publications referring to "big data," including those in the healthcare field and across all fields. Despite the fact that big data has recently gained popularity, the issues at the root of the problem have been around for a long time and have been actively pursued in health research. Meaningful datasets that are too big, too fast, and too complicated for healthcare providers to process and interpret with existing tools are the focus of big data in health [1, 2].

Given the demands of a population that is constantly expanding and has an inverted age pyramid, as well as the paradigm shift in the delivery of health services toward prevention, it is driven by ongoing efforts to make health services more efficient and sustainable, early intervention, and optimal management. For instance, by gathering and investigating influenza related catchphrase looks, Google has built up the Flu Trends administration to distinguish provincial influenza episodes in close ongoing. In particular, Google Flu Trends gathers verifiable inquiry recurrence information of 50 million regular catchphrases in every week from 2003 to 2008. At that point a direct model is utilized to figure the connection coefficient between every watchword seek history information and the real flu like sickness history information acquired from the Centers for Disease Control and Prevention (CDC) in the US [3]. From that point forward, the catchphrases with the most elevated connection coefficients are chosen and their moment seek frequencies are collected to anticipate future influenza flare-ups in the US. With enormous information in catchphrase seeks, Google Flu Trends can recognize influenza episodes over a week sooner than CDC, which can fundamentally lessen the misfortune brought on by this season's cold virus and even spare lives. Another illustration originates from the United Parcel Service (UPS), who outfits its vehicles with sensors to track their speed and area. With the detected information, UPS has advanced its conveyance courses and cut its fuel utilization by 8.4 million gallons in 2011.3 It has been accounted for that enormous information examination is among the main 5 impetuses that assistance with expanding US efficiency and bringing the GDP up in the coming years [4, 5].

II. HEALTHCARE SYSTEM

For the healthcare industry, cloud and big data not only are important techniques but also are gradually becoming the trend in healthcare innovation. Nowadays, medicine is relying much more on specific data collection and analysis, whereas medical knowledge is explosively growing [6]. Therefore, medical knowledge published and shared via cloud is popular in practice. Patients typically will know more than a doctor. As such, the information and knowledge base can be enriched and shared by the doctors over the cloud. The patients can also actively participate in medical activities assisted by big data. Through smart phones, cloud computing, 3- D printing, gene sequencing, and wireless sensors, the medical right returns to the patients, and the role of a doctor is as a consultant to provide decision support to the patients. The revolutions of cloud and big data may have a strong impact on the healthcare industry, which has even been reformed as a new complex ecosystem [5]. Although the innovations are in the healthcare field, there are some issues that need to be solved, particularly the heterogeneous data fusion and the open platform for data access and analysis. For example, although studies are focused on the interconnection between body area networks (BANs)and the cooperation between BANs and medical institutions, it is difficult to fuse the multisource heterogeneous data and the corresponding managements without unified standards and systems. Thus, the healthcare data stored together on the physical layer are still logically separated [7, 8].

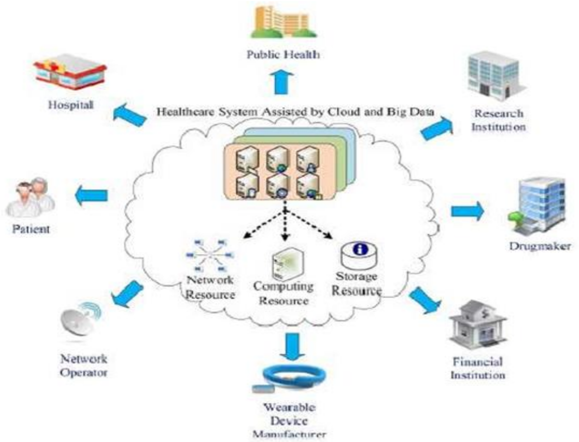


Figure 1: Illustration for the extended healthcare ecosystem

III. MACHINE LEARNING

The main property of an ML is its capability to learn. Learning or preparing is a procedure by methods for which a neural system adjusts to a boost by making legitimate parameter modifications, bringing about the generation of wanted reaction. Learning in an ML is chiefly ordered into two classes as [9].

- Supervised learning
- Unsupervised learning

Supervised Learning

Regulated learning is two stage forms, in the initial step: a model is fabricated depicting a foreordained arrangement of information classes or ideas. The model developed by investigating database tuples portrayed by traits. Each tuple is expected to have a place with a predefined class, as dictated by one of the qualities, called to have a place with a reclassified class, as controlled by one of the traits called the class name characteristic. The information tuple are dissected to fabricate the model all things considered from the preparation dataset.

Unsupervised learning

It is the kind of learning in which the class mark of each preparation test isn't knows, and the number or set of classes to be scholarly may not be known ahead of time. The prerequisite for having a named reaction variable in preparing information from the administered learning system may not be fulfilled in a few circumstances.

Data mining field is a highly efficient techniques like association rule learning. Data mining performs the interesting machine-learning algorithms like inductive-rule learning with the construction of decision trees to development of large databases process. Data mining techniques are employed in large interesting organizations and data investigations. Many data mining approaches use classification related methods for identification of useful information from continuous data streams.

Nearest Neighbors Algorithm

The Nearest Neighbor (NN) rule differentiates the classification of unknown data point because of closest neighbor whose class is known. The nearest neighbor is calculated based on estimation of k that represents how many nearest neighbors are taken to characterize the data point class. It utilizes more than one closest neighbor to find out the class where the given data point belong termed as KNN. The data samples are required in memory at run time called as memory-based technique. The training points are allocated weights based on their distances from the sample data point. However, the computational complexity and memory requirements remained key issue. For addressing the memory utilization problem, size of data gets minimized. The repeated patterns without additional data are removed from the training data set.

Naive Bayes Classifier

Naive Bayes Classifier technique is functioned based on Bayesian theorem. The designed technique is used when dimensionality of input is high. Bayesian Classifier is used for computing the possible output depending on the input. It is feasible to add new raw data at runtime. A Naive Bayes classifier represents presence (or absence) of a feature (attribute) of class that is unrelated to presence (or absence) of any other feature when class variable is known. Naïve Bayesian Classification Algorithm was introduced by Shinde S.B and Amrit Priyadarshi (2015) that denotes statistical method and supervised learning method for classification. Naive Bayesian Algorithm is used to predict the heart disease. Raw hospital dataset is employed. After that, the data gets preprocessed and transformed. Finally by using the designed data mining algorithm, heart disease was predicted and accuracy was computed.

Support Vector Machine

SVM are used in many applications like medical, military for classification purpose. SVM are employed for classification, regression or ranking function. SVM depends on statistical learning theory and structural risk minimization principal. SVM determines the location of decision boundaries called hyper plane for optimal separation of classes as described in figure 1.4. Margin maximization through creating largest distance between separating hyper plane and instances on either side are employed to minimize upper bound on expected generalization error. Classification accuracy of SVM not depends on dimension of classified entities. The data analysis in SVM is based on convex quadratic programming. It is expensive as quadratic programming methods need large matrix operations and time consuming numerical computations.

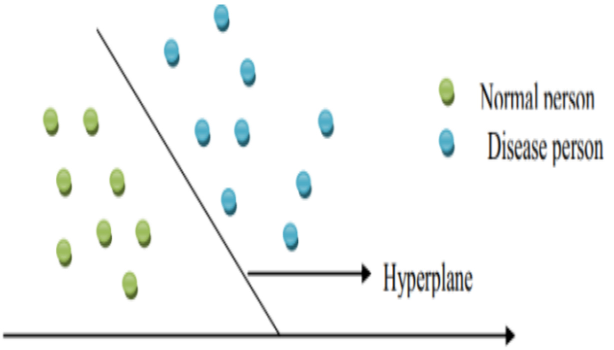


Figure 2: Support Vector Classification

IV. PROPSOED METHODOLOGY

Feature optimization is an important area of health care domain. The extraction process gives the better amount of feature for the feature for the processing of feature. But the signal image generate huge amount of feature for the processing of optimization. In this section used feature optimization technique. The feature optimization technique adopted the particle of swarm optimization.

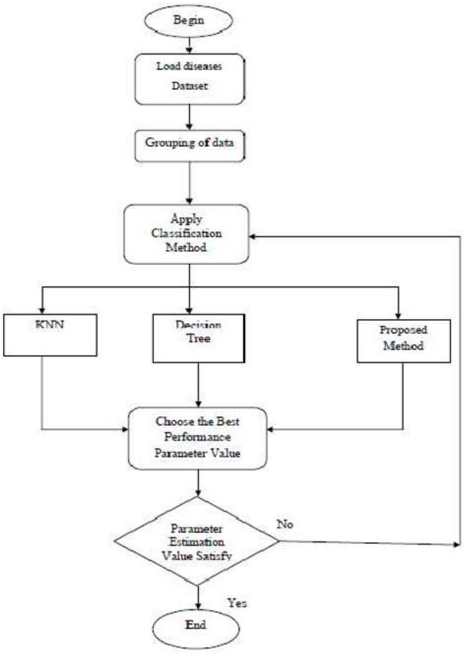

Begin
Load diseases Dataset
Grouping of data
Apply Classification Method
KNN
Decision Tree
Proposed Method
Choose the Best Performance Parameter Value
Parameter Estimation Value Satisfy
No
Yes
End

Figure 3: Proposed model for the diseases diagnosis system

```
begin t=0;
initialize particles p (t); evaluate particles p (t);
while (termination conditions are unsatisfied) begin
t=t+1;
update weights
select pbest for each particl111e select gbest from p (t-1);
calculate particle velocity p(t) calculate particle position
p(t) evaluate particles p(t)
end
end
```

Step 1- begin the process of Health care medical science system using upload the UCI dataset.
Step 2- After the successful UCI dataset uploading process we apply the K-means clustering methods for the arrange the data in a grouping for each respective dataset.
Step 3- after the successfully formation of group we apply the classification methods for each dataset.
Step 4- Apply the classification techniques for the selected dataset such as decision tree, k nearest neighbor classification and proposed method.
Step 5- the proposed methods used with classification and optimization methods.
Step 6- Compute the fitness functions using particle of swarm optimization and update the velocity and position for each particle.
Step 7- Select the best optimal features.
Step 8- we getting the some performance parameters value after applying the Classification and optimization techniques i.e. accuracy, if not is good repeat step 3 to step 7.
Step 9- finally we compare the all performance parameters value and we found that our Proposed methods gives us better results than other methods.
Step 10- Exit the experimental simulation process.

V. EXPERIMENTL RESULTS

Diabetes Dataset

Diabetes patient records were obtained from two sources: an automatic electronic recording device and paper records.

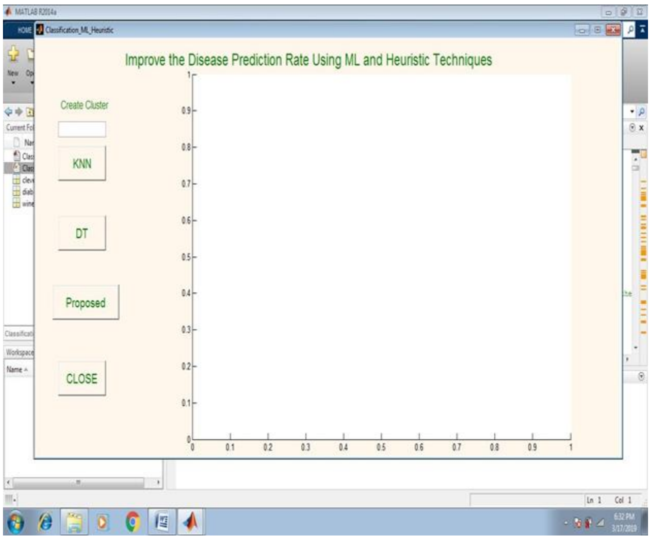
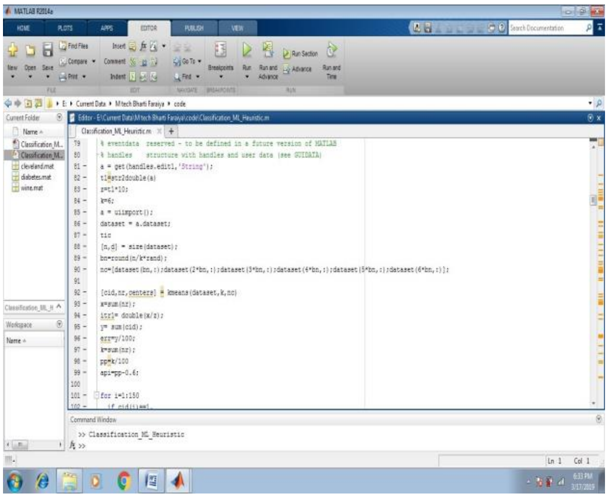

Improve the Disease Prediction Rate Using ML and Heuristic Techniques

Figure 4: Our simulation environment

The automatic device had an internal clock to timestamp events, whereas the paper records only provided "logical time" slots (breakfast, lunch, dinner, bedtime). For paper records, fixed times were assigned to breakfast (08:00), lunch (12:00), dinner (18:00), and bedtime (22:00). Thus paper records have fictitious uniform recording times whereas electronic records have more realistic time stamps.



```
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
a = get(handles.edit1,'String');
t1=str2double(a)
z=t1*10;
k=6;
a = uiimport();
dataset = a.dataset;
tic
[n,d] = size(dataset);
bn=round(n/k*rand);
nc=[dataset(bn,:);dataset(2*bn,:);dataset(3*bn,:);dataset(4*bn,:);dataset(5*bn,:);dataset(6*bn,:)];

[cid,nr,centers] = kmeans(dataset,k,nc)
x=sum(nr);
itr1= double(x/z);
y= sum(cid);
err=y/100;
k=sum(nr);
pp=k/100
api=pp-0.6;

for i=1:150
  if cid(i)==1,
```

```
>> Classification_ML_Heuristic
>>
```

Figure 5: our simulation code environment.

Table 1: comparative result analysis studies

Dataset Name	Method	Accuracy (%)	Precision (%)	Recall (%)
Diabetes	KNN	89	90	91
	Decision Three	94	92	92
	Proposed	95	97	96

VI. CONCLUSION

It is not uncommon that the healthcare data contains biases, noise, and abnormalities, which poses a potential threat to proper decision-making processes and treatments to patients.

Data mining is an analytic process that is designed to search and explore large-scale data (big data) to discover consistent and systematic patterns.

One of the main challenges in big data mining in the medical domain is searching through unstructured and structured medical data to find a useful pattern from patients' information.

In this dissertation we focus on pattern extraction and pattern analysis of healthcare data environment using various classification techniques.

The swarm intelligence family method such as particle swarm optimization is used with the classification techniques and calculate the performance parameter estimation and find the best optimal value.

REFERENCES

[1] Min chen, yixue hao, kai hwang, lu wang, and lin wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities", special section on healthcare big data, vol-5, ieee, 2017. Pp 8869-8879.

[2] Sudha Ram, Wenli Zhang, Max Williams, Yolande Pengetnze, "Predicting Asthma-Related Emergency Department Visits Using Big Data", ieee journal of biomedical and health informatics, vol. 19, 2015. Pp 1216-1218.

[3] Marco Viceconti, Peter Hunter, and Rod Hose, "Big Data, Big Knowledge: Big Data for Personalized Healthcare", ieee journal of biomedical and health informatics, vol. 19, 2015. Pp 1209-1215.

[4] Javier Andreu-Perez, Carmen C. Y. Poon, Robert D. Merrifield, Stephen T. C. Wong, and Guang-Zhong Yang," Big Data for Health", ieee journal of biomedical and health informatics, vol. 19, 2015. Pp 1193-1206.

[5] Yin Zhang, Meikang Qiu, Chun-Wei Tsai, Mohammad Mehedi Hassan, Atif Alamri, "Health-CPS: Healthcare Cyber-Physical System Assisted by Cloud and Big Data", IEEE SYSTEMS JOURNAL, 2015. Pp 1-9.

[6] Ashwin Belle, Raghuram Thiagarajan, S. M. Reza Soroushmehr, Fatemeh Navidi, Daniel A. Beard, Kayvan Najarian, "Big Data Analytics in Healthcare", Hindawi Publishing Corporation Bio-Medical Research International, 2015. Pp 1-17.

[7] Wullianallur Raghupathi, Viju Raghupathi, "Big data analytics in healthcare: promise and Potential", Raghupathi and Raghupathi Health Information Science and Systems 2014. Pp 1-10.

[8] Michael K. K. Leung, Andrew Delong, Babak Alipanahi, Brendan J. Frey, "Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets", IEEE Vol-104, 2016. Pp 176-197.

[9] Daniele Ravı, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, Guang-Zhong Yang,"Deep Learning for Health Informatics",IEEE journal of biomedical and health informatics, VOL. 21, 2017. Pp 4-21.

[10] Michael j. Paul, abeed sarker, johns. Brownstein, azadeh nikfarjam, matthew scotch, karen l. Smith,graciela gonzalez, "social media mining for public health monitoring and surveillance", pacific symposium on biocomputing 2016. Pp 468-477.