IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-4 E-ISSN: 2347-2693

An Improved Disease Prediction System Using Machine Learning

Ajay Kumar [1*], Kamaleshwar M[2], Sanjay Kumar K[3], Sanjith Kumaar R S[4], Arunnehru J[5]

[1]Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India
[2]Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India
[3]Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India
[4]Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India
[5]Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India

[*]*Corresponding Author: sanjaykumarkannan@gmail.com, Tel.: 9600703320*

Available online at: www.ijcseonline.org

Received: 20/Mar/2018, Revised: 28/Mar/2018, Accepted: 19/Apr/2018, Published: 30/Apr/2018

Abstract— There are lots of disease evolving currently due to change in lifestyle, food habits and sleeping habits and there is a lack of technology to identity these. Disease identification using manual checkups is an accurate way but it consumes a lot of time so we need an alternative that performs diseases diagnosis quick and accurate, this leads to need for data analytics and machine learning. Data analytics we analyze the user data and provide insights to the user. We use machine learning techniques to analyze user data and supervised algorithm such as SVM and unsupervised algorithm such as K-Means clustering are used for classification of the datasets .Random forest is used to create decision trees using user data and important data can be extracted from the decision tree.

Keywords— Support vector machine (SVM), Random Forest(RF).

I. INTRODUCTION

Data mining is the action finding out the undiscovered patterns in the datasets .With the help of these patterns we can extract knowledge and we can present it in a form understandable to humans. Data mining plays an important role in medical industry by helping in disease prediction. Some the diseases that can be predicted with the help of data-mining techniques are heart disease, lung cancer, breast cancer etc. Data mining is intensively used in medical domain. By using data mining in analyzes medical industry we can find the unidentified patterns in the datasets. We can perform medical analysis in raw medical data with the help of these patterns. The major cause of casualties in the world is due to heart disease. Most of the deaths occur in countries like India, USA are due to cardio-vascular diseases. Data mining techniques such as Clustering, SVM-polynomial, SVM-RBF, RF, K-NN Algorithms agreed to analyze the different kinds of lung based problems.

II. RELATED WORK

The difficult of recognizing constrained association rules for heart illness prediction was studied by Carlos Ordonez. The data mining techniques have been engaged by various works to analyze various diseases, for instance: Hepatitis, Cancer, Diabetes, Heart diseases. According to WHO (World Health Organization), heart disease is the main cause of death in the UK, USA, Canada, England [2]. Heart disease kills one in every 32 seconds in USA .25.4% of all deaths in the USA today are caused by heart disease. Classification is one of the supervised learning methods to extract models describing important classes of data. Three classifers used to diagnose disease and suggest prediction are Support Vector Machine(SVM),K Means Clustering and Random Forest.

III.METHODOLOGY

A.Support Vector Machine

The basic idea behind SVM application in pattern recognition is to use a hyperplane as decision plane, which not only separates the two-class samples but also maximizes the recognition margin, as shown in figure 1. The optimization problem is ultimately transferred into a convex quadratic problem. Here, ω is normal vector of the hyperplane; the distance is determined by the γ.

Given a training sample $\{(x_i, y_i), i = 1, 2, l\}$ with size of l, it has two types: If $x_i \in R^n$ belongs to type 1, it would be written as positive $(y_i = 1)$; If it belongs to type 2, it would be written as negative $(y_i = -1)$. The learning objective is to construct a decision function to correctly classify the test samples. $x'\omega - \gamma = \pm 1$ represents the classification hyperplane with two different classes, the sign of data is determined by the following equations:

$$x_i'\omega - \gamma \geq \pm 1 \qquad y_i = 1$$

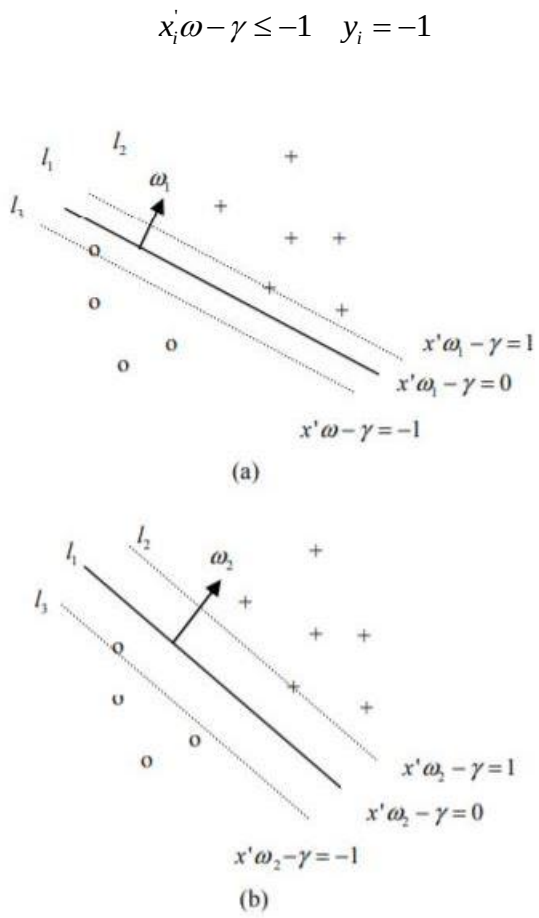

Fig 1. Maximized Margin Method

B. Radial Basis Function Neural Network

Radial Basis Function Neural Network Radial Basis Function Network or RBFN is a particular type of neural network. And it is generally is used as a non-linear classifier. When people talk about Artificial Neural Networks it mostly refers to Multilayer Perceptron. In MLP the input of every neuron is the weighted sum of its inputs. The sum is obtained by multiplying each input value by a coefficient. A simple linear classifier is a single MLP neuron , by combining these neurons into a network we can build complex non-linear classifiers. Classification is done by RBFN by measuring the similarity of the input compared with the training dataset. So RBFN is a better approach then MLP[5]. RBFN have many neuron and each of them stores a prototype, which is an example obtained from the training set[3]. The euclidean distance between the input and its prototype is calculated for every neuron, if we want to classify a new input. Roughly, if the input more closely resembles one of the two classes A and B, input is classified to the more similar class.

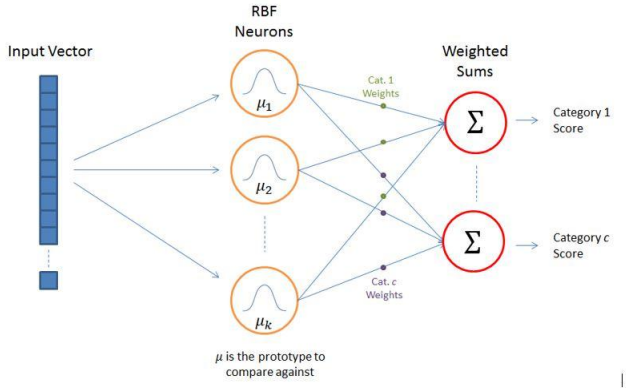

Fig 2. RBF input and output

Architecture above is pictorial representation of a RBF network. The architecture consists of a layer of RBF neurons, input vector and an output layer with a class of data.

The Input Vector: The input vector which we are trying to classify is a n-dimensional vector. All the input vectors are linked to each of the RBF neurons.

The RBF Neurons: A model vector is put away inside each RBF neuron which is only one of the vectors from the training set. A random value between 0 to 1 is generated when RBF neuron compares its prototype with the input vector and that is used to measure similarity. If the input value and prototype values are equal, output generated from that RBF neuron will 1.When the distance between the input and prototype grows, the response falls drastically towards 0.RBF neuron is in the form of a bell curve like indicated in the architecture diagram. Response value of neuron is called the activation value[4]. Neurons centre is often called as the prototype vector because the value is present in the centre of a bell curve.

C. K-Mean clustering

K-means clustering is an unsupervised learning and it is utilized when you have unlabelled information. The objective of this calculation is to nd clusters in the information, with the quantity of gatherings spoke to by the variable K. The formula works iteratively to assign a data point to one of K clusters in view of the highlights that are given. One of the famous K-means algorithm is centroid based clustering algorithm .

In order to execute a k-means algorithm we have to assign three points called cluster centroids. For grouping our data into three clusters we need to have three cluster centroids. Two steps of K-means iterative algorithm is: 1. Cluster initializing 2. Move centroid . Cluster assignment is random, based on the number of cluster centroid assigned. In the next step, the movement of cluster centroid is determined by density of data points in the graph, i.e the cluster centroids tends to move to areas which are densely populated[6]. Clusters are created when the centroid remains stationary.

This occurs when the proximity between the cluster centroid and data points is minimum.

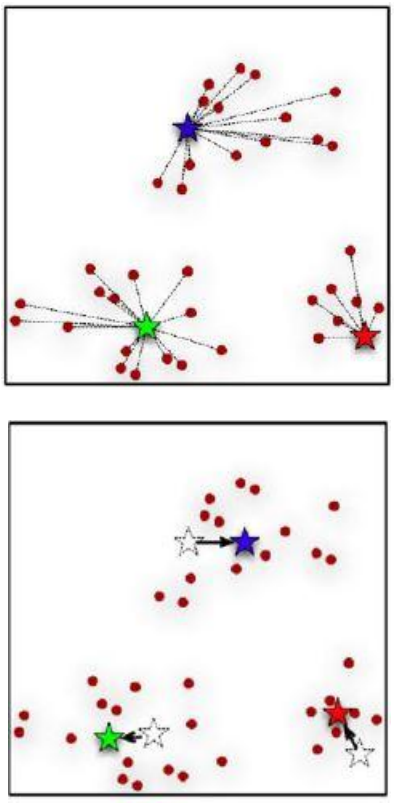

Fig 3. Move centroid step in K means algorithm

D. Random Forest Algorithm

Random Forest algorithm is a supervised learning classification algorithm and it creates a forest of trees with the help of input. Usually forest looks robust when there are more number of trees present in it[7]. Similarly using random forest classifier we can increase accuracy with increasing number of trees in the forest. By knowing decision tree algorithm, we can create more decision tees .The calculation on how to select the node will be same for same dataset. Assumptions on creating Decision Tree: At the outset the whole preparing set is considered as a root. The esteems highlighted are wanted to be straight out.

In the event that the qualities are continuous then they are discretized before building the model. Depending upon the attribute values the records are repeated. For placing attributes as root or internal nodes of the tree a statistical approach is followed.

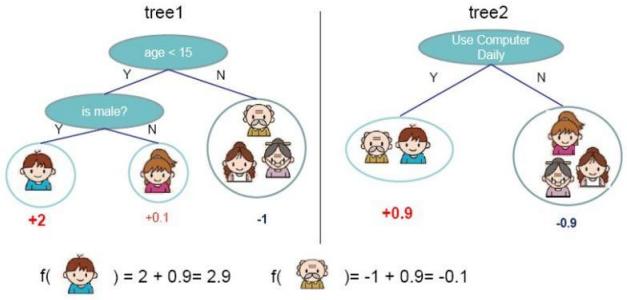

Fig 4. An example of decision tree mode

E. Basic decision tree concept

Rule based system mostly used decision tree concept. The training dataset has targets and features, some rules will be generated by the decision tree algorithm[8]. The prediction on the test dataset is done with the help of same test rules. In decision tree algorithm forming rules and calculating these nodes will happen using gain and gini index calculations. In RF algorithm, rather than using information gini or gain index for evaluating the root node, splitting the feature nodes and the action of finding the root node will happen randomly.

Fertility: Fertility is the characteristic ability to deliver offspring. By estimating fertility rate is the checking of offspring conceived per mating pair, it might be individual or population. Fertility varies from fecundity, which is characterized as the potential for generation . Absence of fecundity, while absence of fertility is infertility would be called sterility. Human fertility relies upon elements of sustenance, sexual conduct, connection, culture, nature, endocrinology, timing, financial aspects, lifestyle and feelings.

Drawback: Infertility happens after having regular unprotected sex when a couple cannot conceive[9]. It might be that one partner can't add to origination, or that a woman can't convey a pregnancy to full term. Without the utilization of anti-conceiving medication characterized as not considering following a year of sexual intercourse. In the United States, around 10 percent of women's matured 15 to 44 years are assessed to experience issues considering or remaining pregnant. The issue experienced by 8 to 12 percent of couples by fertility in around the world. In the vicinity of 45 and 50 percent of cases are thought to come from factors that influence the man.

Lung Cancer: In one or both lungs that start off the uncontrolled growth of abnormal cells by lung cancer. Healthy lung tissue does not develop in the abnormal cells, they divide rapidly and form tumours. They undermine the lung's ability to provide the bloodstream with oxygen as tumours become larger and more numerous. Don't spread

and stay set up that tumours are known as ”benign tumours”. The most unsafe one is Malignant tumours, spread through the circulatory system or the lymphatic framework either to different parts of the body. Metastasis alludes to cancer spreading past Its site of beginning to different parts of the body will spread disease called as ”Metastasis”. It is considerably harder to treat effectively when disease spreads. Primary lung cancer begins in the lungs, while optional lung tumour begins elsewhere in the body, metastasizes, and achieves the lungs. They are viewed as various sorts of cancers and are not treated similarly. As per the National Cancer Institute, before the finish of 2015, there will have been 221,200 new lung cancer analyse and 158,040 lung-malignancy related passing's in the USA. As indicated by the World Health Organization or WHO, Cancer was caused 7.6 million passing in every year and represents to 13 of every worldwide demise. As observed underneath, lung malignancy is by a wide margin the main disease executioner.

IV. EXPERIMENTAL RESULT

PRECISION

It is the fraction of relevant information returned of all retrieved information. The precision is the proportion tp/(tp + fp) where tp is the quantity of true positives and fp the quantity of false positives. The accuracy is naturally the capacity of the classifier not to name as positive a sample that is negative.

RECALL

It is the fraction of relevant information returned of all relevant information. The recall is the proportion tp/(tp + fn) where tp is the quantity of true positives and fn the quantity of false negatives[10]. The recall is naturally the capacity of the classifier to discover all the positive samples.

F-Measure

F-measure is a metric of a test's accuracy. It considers both the precision p and the recall r of the test to figure the score. The F1 score is the symphonious normal of the precision and recall, where a F1 score achieves its best an incentive at 1 and most noticeably awful at 0. The F-beta score can be translated as a weighted consonant mean of the precision and recall, where a F-beta score achieves its best score at 1 and pessimistic standpoint score at 0.

DATASET DESCRIPTION

Lung Cancer dataset

The following are the generic lung cancer attributes used Coughing up blood (heamoptysis) or bloody mucus, Weight loss and loss of appetite, Wheezing, Shortness of breath, Fatigue and weakness, Swelling of the neck and face, Fever, Loss of appetite, Nausea and vomiting. We use all these attributes to determine the possibility of disease in a human being.

Fertility dataset

Some of the attributes that are generally considered for determining disease in a person are: Season in which the analysis was performed, Age at the time of analysis, Childish diseases (ie , chicken pox, measles, mumps, polio),Accident or serious trauma ,Surgical intervention ,High fevers in the last year ,Frequency of alcohol consumption, Smoking habit, Number of hours spent sitting per day[11].

Parkinsons dataset

Matrix column entries that are used to identify the presence of disease in a human being are name - ASCII subject name and recording number, MDVP(Fo(Hz)) - Average vocal major frequency, MDVP(Fhi(Hz)) - Maximum vocal essential recurrence, MDVP(Flo(Hz))-Minimum vocal basic frequency, Jitter(Abs)- Several measures of variety in key frequency, Shimmer(DDA) - Several measures of variety in amplitude, NHR, HNR - Two measures of proportion of commotion to tonal segments in the voice ,status - Health status of the subject (one) Parkinson's, (zero) - healthy,D2 - Two nonlinear dynamical many-sided quality measures, DFA - Signal fractal scaling type, PPE - Three nonlinear measures of principal recurrence variety.

Confusion Matrix

True positives (TP) When the result predicted is 'yes' and the actual result is also true then it is called true positive.

True negatives (TN)

When the result predicted is 'no' and the actual information is also 'no' then it is called as true negative.

False positives (FP)

When the result predicted is 'yes' but in reality it is actually 'no' then it is called as false positive. This leads to type 1 error.

False negatives (FN)

When the result predicted is 'no' but in reality it is actually 'yes' then it is called as false negative. This leads to type 2 error.

Correctly Classified Instances	83
Incorrectly Classified Instances	17
Kappa statistic	0.6072
Mean absolute error	0.2785
Root Mean Squared Error	0.3582
Relative Absolute Error	61.8292
Root Relative Squared Error	75.5279
Total Number Of Instances	100

TP Rate	FP Rate	Precision	Recall	F-Measure
0.909	0.324	0.845	0.909	0.876
0.676	0.091	0.793	0.676	0.730
0.830	0.224	0.827	0.830	0.826

Figure 5. Performance Measure for Fertility Dataset

Correctly Classified Instances	63
Incorrectly Classified Instances	37
Kappa statistic	0.2492
Mean absolute error	0.4764
Root Mean Squared Error	0.486
Relative Absolute Error	95.5292
Root Relative Squared Error	97.288
Total Number Of Instances	100

TP Rate	FP Rate	Precision	Recall	F-Measure
0.736	0.489	0.629	0.736	0.678
0.511	0.264	0.632	0.511	0.565
0.630	0.384	0.630	0.630	0.625

Figure 6. ***Performance Measure for Lung Dataset***

Correctly Classified Instances	84
Incorrectly Classified Instances	16
Kappa statistic	0.0643
Mean absolute error	0.2409
Root Mean Squared Error	0.3375
Relative Absolute Error	92.345
Root Relative Squared Error	94.2829
Total Number Of Instances	100

TP Rate	FP Rate	Precision	Recall	F-Measure
0.976	0.933	0.856	0.976	0.912
0.067	0.024	0.333	0.067	0.111
0.840	0.797	0.777	0.840	0.792

Figure 7. ***Performance Measure for Parkinsons Dataset***

VI. Conclusion and Future Scope

In this paper we have presented an efficient approach for fragmenting and extracting substantial forms from the trained data warehouses for the efficient prediction of disease analysis using various algorithms. We have decided to integrate the analytics module to the equipment used to perform checkups on human beings so that no need of manual entering of data, and the prediction is displayed in the monitor linked with the test equipment. And also future work involves adding all the diseases that are currently present as well as dead diseases to the system, gathering dataset for all datasets and create a solid prediction model. By using SVM suitable accuracy is achieved and the classification is clear and the model can be predict without any flaw as we have attained good accuracy. We were able to achieve good precision, recall with current dataset, we are trying to achieve more accuracy by accumulating more dataset from all possible sources.

VII. References

[1] V. Manikantan and S. Latha, Predicting the analysis of heart disease symptoms using medicinal data mining methods, International Journal of Advanced Computer Theroy and Engineering, vol.2, pp.46-51,2013.

[2] Yuh-Jye Lee and O.L. Mangasarian. SSVM: A smooth support vector machine. Technical Report 99-03, Data Mining Institute, Computer Science Department, University of Wisconsin, Madison, Wisconsin, September 1999.

ComputationalOptimizationandApplications20(1), October 2001.

[3] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni Predictive data mining for medical diagnosis: an overview of heart disease prediction International Journal of Computer Science and Engineering, vol. 3 ,2011.

[4] R. Agrawal,T Imielinski ,and A. Swami ,Mining association rules between sets of items in large databases.

[5] Hnin Wint Khaing, Data Mining based Fragmentation and Prediction of Medical Data, International Conference on Computer Research and Development, ISBN: 978-1-61284-8402,2011.

[6] M. Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar, Enhanced prediction of heart disease with feature subset selection using genetic algorithm, International Journal of Engineering Science and Technology vol.2, pp.5370- 5376,2010.

[7] Douglas Burdick, Manuel Calimlim, Johanne Gehrke,MAFIA: A Maximal Frequent Item set Algorithm For Transactional Databases, Proceedings of the 17th International Conference on Data Engineering.

[8] S.Vijayarani, M. Divya, An Efficient Algorithm for Generating Classification Rules, IJCST ,vol. 2, Issue 4, 2011.

[9] M.C. Ferris and T.S. Munson. Interior point methods for massive support vector machines. Technical Report 00-05, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, May 2000.

[10] M. Brown, W.Grundy, N. Cristianini D. Lin, C. Sugnet, T.Furey, M. Ares Jr., and D. Haussler. Knowledge-based analysis of microarray gene expression data using support vector machines. Proceedings of the National Ac.

[11] J. e. Dennis and R. B. Schnabel. Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Prentice-Hall, Englewood Cliffs, N. J., 1983.

