Disease Prediction by Using KNN Algorithm over Big Data

B.Vani[1], Mr. G. Ananthnath [2]
[1]Student, Dept of MCA, KMM Institute of Post-graduation Studies
[2]Asst. Professor, Dept of MCA, KMM Institute of Post-graduation Studies

Abstract- **In the present days diseases are leading the chart of causes of deaths in human beings , along with road accidents and suicides. The prevention of road acciedents and suicides depends upon the victims individual mental and physical abilities. Hence no body can predict the happening of these two : But in the case of diseases, we can predict the occurance in future depending upon the present health condition of the individuals. Many of the rural areas do not have hospitals and the physicians are not available to do complex tests and predict the diseases. In such situations the disease can be efficiently predicted by observoing and analysing the patients health parameters. The Data Mining techniques can be utilized to do and improve this prediction. Since Data Mining offers many algorithms to implement like Naïve bayes, decision tree and k- nearest neighbour. The present study deals with the usage of KNN algorithm with least and well defined parameters for improving the disease predition. This provides an oppurtunity for easy diognisis of the patients in hospitals and even at home.**

Index Terms- **Data mining techniques, diease prediction, KNN algorithm, Naive Bayes, Decision Tree, Machine to Machine technology**

I. INTRODUCTION

In the recent generations, the usage of computer is increased rapidly. Since the usage of computers at work places can reduce the time, effort and number of employees the digital technology is being implement every where from marine to space. In order to satisfy a wide domain of end users a wide variety of softwares has to be developed. This offers huge number of employements which can cause the living standants besides those advantages, the computerization may reduce the physical strain but, leads to lot of mental strain among the software developers sitting for long times infront of computers. Due to long sitting and lack of physical strain the stress and cholestrol levels are increased and leads to blood pressure fluxuations, heart diseases and diabetes. As per the survey % of population are addicted to net surfing. Not only the computerization, the individual living habbits like drinking, alchohal, smoking, consumption of norcotic drugs and lack of physical exersises are also leading to the kind of diseases. In certain cases heridity can also cause for this diseases.

However , the same software technology can be utilized to predict these diseases efficiently. Researches are working data mining technologies for the effective presdiction of a wide range of diseases by processing the enormous amount of data available in the hospitals. The prediction using data mining techniques are not only limitted to health care but it was spread into a wide variety of sectors like weather forcasting [1] and stock price prediction [2] etc. Here, some basic introduction about the commonly used algorithms is given below

II . ALGORITHMS

There are some algorithms that have been used for this purpose like Naive Bayes, Decision Tree, and k-Nearest Neighbor (KNN).Here in this paper we have used KNN algorithm .

A . Naive Bayes:

The Naive Bayes algorithm represents a supervised machine learning method for classification. It uses a probabilistic model by determining probabilities of the outcomes. It is used in diagnostic and predictive problems. Naive Bayes is robust to noise in input dataset.

B . Decision Tree:

Decision tree learning uses a decision tree as a predictive model which maps input about an item to output of the item. Tree models with finite classes of output are called classification trees. In these tree structures, leaves represent class labels and branches represent relation between attributes that results in those class labels. Decision trees with continuous output classes are called regression trees. In data mining, a decision tree can be an input for decision making.

C . KNN- Algorithm:

In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

III . PROCEDURE

- In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.
- In k-NN regression, the output is the property value for the object. This value is the average of the values of its *k* nearest neighbors.

The algorithm on how to compute the K-nearest neighbors is as follows:

1. Determine the parameter K = number of nearest neighbors beforehand. This value is all up to you.
2. Calculate the distance between the query-instance and all the training samples. You can use any distance algorithm.
3. Sort the distances for all the training samples and determine the nearest neighbor based on the K-th minimum distance.
4. Since this is supervised learning, get all the Categories of your training data for the sorted value which fall under K.
5. Use the majority of nearest neighbors as the prediction value.

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

Both for classification and regression, a useful technique can be to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/*d*, where *d* is the distance to the neighbor.

The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

A peculiarity of the *k*-NN algorithm is that it is sensitive to the local structure of the data. The algorithm is not to be confused with k-means, another popular machine learning technique.

The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples.

In the classification phase, *k* is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the *k* training samples nearest to that query point.

A commonly used distance metric for continuous variables is Euclidean distance. For discrete variables, such as for text classification, another metric can be used, such as the overlap metric (or Hamming distance). In the context of gene expression microarray data, for example, *k*-NN has also been employed with correlation coefficients such as Pearson and Spearman. Often, the classification accuracy of k-NN can be improved significantly if the distance metric is learned with specialized algorithms such as Large Margin Nearest Neighbor or Neighborhood components analysis.

A drawback of the basic "majority voting" classification occurs when the class distribution is skewed. That is, examples of a more frequent class tend to dominate the prediction of the new example, because they tend to be common among the *k*nearest neighbors due to their large number. One way to overcome this problem is to weight the classification, taking into account the distance from the test point to each of its *k* nearest neighbors. The class (or value, in regression problems) of each of the *k* nearest points is multiplied by a weight proportional to the inverse of the distance from that point to the test point. Another way to overcome skew is by abstraction in data representation. For example, in a self-organizing map (SOM), each node is a representative (a center) of a cluster of similar points, regardless of their density in the original training data. *K*-NN can then be applied to the SOM.

Let m be the number of training data samples. Let p be an unknown point.

Step-1 Store the training samples in an array of data points arr[]. This means each element of this array represents a tuple (x, y).

Step-2 for i=0 to m:

Step-3 Calculate Euclidean distance d(arr[i], p).

Step-4 Make set S of K smallest distances obtained. Each of these distances correspond to an already classified data point.

Step-5 Return the majority label among S.

IV . CONCLUSION

In this paper, With help of data mining techniques, disease prediction can be improved. There are some algorithms that have been used for this purpose like Naive Bayes, Decision Tree, and k-Nearest Neighbor (KNN). Here KNN algorithm is used. By using this algorithm we can improve accuracy and efficiency comparing Naive Bayes and Decision tree.

REFERENCES

[1] Srinivas, K., “Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques”, IEEE Transaction on Computer Science and Education (ICCSE), p(1344 - 1349), 2010.

[2] Shanta kumar, B.Patil,Y.S.Kumaraswamy, “Predictive data mining for medical diagnosis of heart disease prediction” IJCSE Vol .17, 2011

[3] M. Anbarasi et. al. “Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm”, International Journal of Engineering Science and Technology Vol. 2(10), 5370-5376 ,2010.

[4] Hnin Wint Khaing, “Data Mining based Fragmentation and Prediction of Medical Data”, IEEE, 2011.

[5] Srinivas, K.,” Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques”, IEEE Transaction on Computer Science and Education (ICCSE), p(1344 - 1349), 2010.

[6] Yanwei Xing, “Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease”, IEEE Transactions on Convergence Information Technology, pp(868 – 872), 21-23 Nov. 2007

[7] IBM, Data mining techniques, http://www.ibm.com/developerworks/opensource/library/ba-data-miningtechniques/ index.html?ca=drs- , downloaded on 04 April 2013.

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

[9] Shadab Adam Pattekari and Alma Parveen,“Prediction system for heart disease using Naïve Bayes”, International Journal of Advanced Computer and Mathematical Sciences, vol.3,pp 290-294,2012.

[10] Jyoti Soni, Ujma Ansari, Dipesh Sharma and Sunita Soni, “Predictive data mining for medical diagnosis: an overview of heart disease prediction”, International Journal of Computer Science and Engineering, vol. 3, pp.43-48, 2011.

[11] Hnin Wint Khaing, “Data Mining based fragmentation and prediction of medical data”, International Conference on Computer Research and Development, ISBN: 978-1-61284-840-2, 2011.

[12] K.Shekar, N.Deepika and D.Sujatha,“Association rule for classification of heart-attack patients”, International Journal of Advanced Engineering Sciences and Technologies, vol.11, no. 2, pp.253-257, 2011.

[13] M. Anbarasi, E. Anupriya and 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.

[14] N.A. Setiawan, P.A. Venkatachalam, and Ahmad Fadzil M.H. , Rule Selection for Coronary Artery Disease Diagnosis Based on Rough Set,

International Journal of Recent Trends in Engineering, Vol 2, No. 5, November 2009.

[15] Sunita Soni , Jyothi Pillai, O.P.Vyas, An Associative Classifier Using Weighted Association Rule , IEEE proceedings of the World Congress on Nature and Biologically Inspired Computing (NaBIC'09), December 09-11, 2009, 1492-1496.

[16] Shantakumar B.Patil, Y.S.Kumaraswamy, Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656

[17] Ruben D. Canlas Jr.,data mining in healthcare: current applications and issues, August 2009.

[18] Sellappan Palaniappan Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008

[19] Asha Rajkumar, G.Sophia Reena, Diagnosis Of Heart Disease Using Datamining Algorithm, Global Journal of Computer Science and Technology 38 Vol. 10 Issue 10 Ver. 1.0 September 2010.

[20] K.Srinivas, B.Kavihta Rani , A.Govrdhan , Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks, (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, 250-255.