International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2017, pp. 366-369
e-ISSN:2278-621X

A COMPARATIVE STUDY ON DATA CLASSIFICATION ALGORITHMS KNN AND SVM IN DIAGNOSING HEART DISEASE

ShilpaJoseph[1], Abdul Ashif D H[2] & Vanitha T[3]

Abstract: Over past 10 years the researches said that the leading reason for deaths in India is Heart diseases. Data mining algorithms applied in healthcare industry play a significant role in prediction and diagnosis of the diseases.Data mining in healthcare sector is a developing field of high significance for giving anticipation and a more profound comprehension to healthcare data. health data mining endeavors to take care of true medical issues in determination and treatment of diseases. Analysts are utilizing data mining methods in the determination of diseases, for example, heart disease,diabetes, stroke and tumor. Many data mining systems are utilized as a part of the conclusion of heart disease indicating distinctive levels of accuracy,memory usage and elapsed time. K-Nearest-Neighbor(KNN) and Support Vector Machine are some data mining techniques used in classification problems in healthcare sector. But, these algorithms are less used in the diagnosis of heart disease patients.This paper investigates applying KNN and SVM to help healthcare professionals in the diagnosis of heart disease.The results show that applying SVM could achieve higher accuracy and efficiency than applying KNN in heart related data sets classification.
Keywords –KNN, SVM, Heart Diagnosis, Data Mining,Vapnik-Chervonenkis.

1. INTRODUCTION

World health organization declared many years back that the increase in the mortality rate is mainly due to the cardiac problems.When it comes to heart disease time is very important and crucial to get correct diagnosis in early stage. Patient having chest pain complaint may undergo unnecessary treatment or admitted in the hospital. But since these data will provide a huge amount of data, that can be used to extract useful information for analyzing and predicting the reasons for the disease, to predict the reasons for the cause, and hence to help the people. Researchers so have been using different data mining techniques to help the healthcare sectors in improvising the betterment process of heart diseases.

Data mining acts as an essential step in the discovery of knowledge.Data mining algorithms which are utilized as a part of health services division gives huge advantages in forecast and conclusion of the diseases. By the utilization of data mining procedures in medicinal services Profitable knowledge can be found. The extensive measure of information produced by health services exchanges are excessively intricate and immense, making it impossible to be processed and analyzed by legacy methods. Data mining gives the technique and innovation to change these hills of data into helpful information for basic leadership.

K-Nearest-Neighbor (KNN) is one among the most used data mining techniques in classification problems and pattern recognition. And Support Vector machine (SVM) modeling is a promising classification approach for determining medication adherence in heart patients. This predictive models stratifies the patients so evidence based choices can be made and patients treated properly.

2. BACKGROUND OF THE STUDY

To extract useful knowledge,health experts store significant amounts of patient's data.Analysts have been examining the utilization of data mining methods and statistical analysisto help medicinal services experts in the finding of coronary illness. Measurable examination has recognized the hazard factors related with heart ailment to be age, pressure, obesity, cholesterol, hypertension, diabetes, family history of heart disease, absence of physical action and habit of smoking. Learning of the hazard factors related with heart infection encourages medical experts to distinguish patients at high danger of having heart disease [2].

Scientists have been applying diverse data mining strategies, such as support vector machine, neural network, naïve Bayes, kernel density, bagging and decision tree over diverse heart diseases datasets to enable wellbeing to mind experts in the finding of heart diseases. The aftereffects of the distinctive data mining research can't be thought about in light of the fact that they have utilized diverse datasets. However, after some time many comparative studies have done with different algorithms and in this paper we have opted KNN and SVM to compare and contrast.

[1] Department of Computer Applications, AIMIT, St. Aloysius college, Mangalore, Karnataka, India
[2] Department of Computer Applications, AIMIT, St. Aloysius college, Mangalore, Karnataka, India
[3] Department of Computer Applications, AIMIT, St. Aloysius college, Mangalore, Karnataka, India

3. COMPARED ALGORITHM

3.1 Support Vector Machine(SVM) algorithm –

Support Vector Machine is a managed machine learning algorithm utilized for both order or relapse challenges. Be that as it may, it is for the most part utilized as a part of order issues. In this algorithm, we plot every datum thing as a point in n-dimensional space with the estimation of each component being the estimation of a specific arrange. At that point, we perform order by finding the hyper-plane that separate the two classes exceptionally well.**S**upport Vector Machine (SVM) is a category of universal feed forward networks like Radial-basis function networks, pioneered by Vapnik. SVM can be used for pattern classification and nonlinear regression. More precisely, the support vector machine is an approximate implementation of the method of structural risk minimization. This guideline depends on the reality the error rate of a learning machine on test data is limited by the sum of the training error rate and term that relies upon the Vapnik-Chervonenkis (VC) measurement.

Optimal Hyperplane for patterns: Consider the training sample $\{(xi, yi)\}_{t=1}^{N}$ where xi is the input pattern for the i^{th} instance and yi is the corresponding target output. With pattern represented by the subset $y_i = +1$ and the pattern represented by the subset $y_i = -1$ are linearly separable. The equation in the form of a hyperplane that does the separation is

$$w^T x + b = 0 \quad (1)$$

where x is an input vector, w is an adjustable weight vector,and b is a bias. Thus,

$$w^T x_i + b \geq 0 \text{ for } y_i = +1 \quad (2)$$

$$w^T x_i + b < 0 \text{ for } y_i = -1 \quad (3)$$

For a given weight vector w and a bias b, the separation between the hyperplane defined in eq. 1 and closest data point is called the margin of separation, denoted by ρ asshown in figure 1, the geometric construction of an optimal hyperplane for a two- dimensional input space [4].

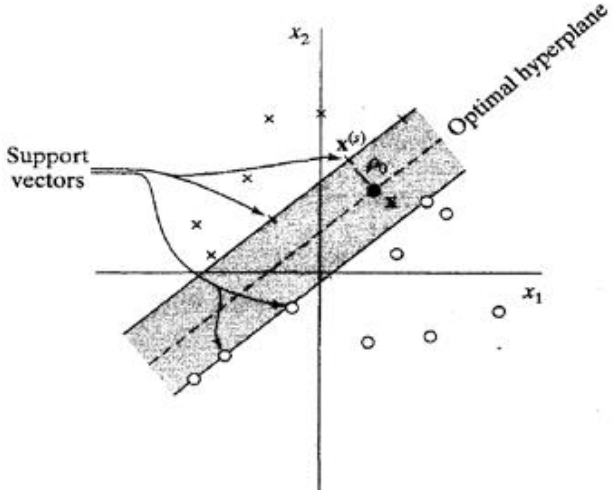

Fig 1. Optimal Hyperplane for a two dimensional input space

The discriminant function gives an algebraic measure of the distance from x to the optimal hyperplane for the optimum values of the weight vector and bias, respectively.

$$g(x) = wo^T x + bo \quad (4)$$

3.2 K Nearest Neighbour (KNN) algorithm –

KNN is one of the simplest and straight forward data mining techniques. K-Nearest-Neighbor is one of the most widely used data mining techniques in classification problems. Its simplicity and relatively high convergence speed make it a popular choice. But, a main disadvantage of KNN classifiers is the requirement of memory it needs to store the whole sample. When the sample is large, response time on a sequential computer is also large. It is showing good performance in classification problems of various datasets despite the memory requirement issue. As the training examples need to be in the memory at run-time it is called Memory-Based Classification. The difference between the attributes is calculated using the Euclidean distance when dealing with continuous attributes. If the first instance is (a1, a2, a3 … an) and the second instance is (b1, b2, b3, … bn), the distance between them is calculated by the following formula:

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2} \quad (5)$$

'The large values frequency swamps the smaller ones' is a major problem when dealing with the Euclidean distance formula [5]. For example, in heart disease records the cholesterol measure ranges between 110 and 200 while the age measure ranges between 40 and 90. So the influence of the cholesterol measure will be higher than the age. We normalize the continuous attributes so that they have the same influence on the distance measure between instances are used, to overcome this problem. Many a times KNN works with continuous attributes however it can work with discrete attributes also. When dealing with

discrete attributes if the attribute values for the two instances a2, b2 are different so the difference between them is equal to one otherwise it is equal to zero [1].

4. EXPERIMENTAL RESULTS

We took 3 different secondary data sets from UCI machine learning data repository, and applied them to the algorithms to find the accuracy ofthe data obtained, Memory used for execution and also the time required for execution i.e., elapsed time. Fig 2 shows the experimental methodology briefly. When the three different data sets are applied we got the results that says that SVM works much better than KNN.

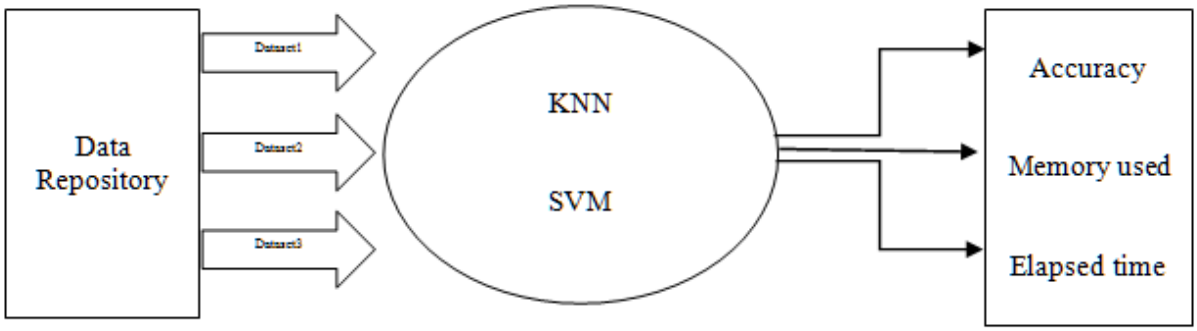

Fig 2. Experiment methodology

Table -1 Experiment Result

	Support Vector Machine			**K Nearest Neighbor**		
	Dataset(1000)	**Dataset(500)**	**Dataset(100)**	**Dataset(1000)**	**Dataset(500)**	**Dataset(100)**
Elapse time(seconds)	3.132	1.518	0.621	9.155	6.471	3.14
Memory used	29337	27210	24665	29249	27636	25267
Accuracy	82.541	76.278	80.63	79.215	76.637	85.345

When the size of the data sets increases KNN loses its accuracy and speed but SVM will maintain its accuracy and memory usage and also the elapsed time. Table 1. above shows the experimental results of KNN and SVM in 3 different criteria i.e., Elapsed time, memory used and Accuracy when we applied the data sets of 1000,500 and 100 data sets in 3 different stages.

- As we can clearly see in the table, elapsed time of KNN is keep high in most of the cases. But as size of data set increases we can see time consumed for train model of both systems increases. but KNN gives more poor results when size of data set increases. It consumes large time for training.
- Amount of main memory with is used to execute the algorithm is defined as memory used which is given in Kilo Bytes. As we can see in the resultant table memory used by both systems are remain constant in all cases.
- As we can clearly see on the table accuracy of K-NN is keep high in most of the cases. But as size of dataset increases we can see accuracy of both system decreases. But K-NN simulates more poor results when size of data set increase.

5. CONCLUSION

Heart disease is the leading cause of death all over the world in the past ten years. Different data mining techniques have been used by researchers with availability of huge amount of patients' data that could be used to extract useful knowledge and hence to help health care analysts in the diagnosis of heart disease to decrease mortality of heart related patients each year.Through this research study after comparing we found that K-NN is a much good classifier but when we apply this algorithm over textual data it's all performance parameters are varies according to the size of dataset. K-NN performs poor results as the size of data set increases it is best fit for small data set. SVM is complex classifier and here we implement leaner kernel. We found that the accuracy and other performance parameters depends over dataset size of the data set and SVM is having good performance in the all three predicted areas than that of KNN.

6. REFERENCES

[1] Hassan Shee Khamis, Kipruto W. Cheruiyot, Stephen Kimani, (2014)"Application of k- Nearest Neighbour Classification in Medical DataMining".

[2] WHO (2009): Primary health care – Now more than ever. World Health Organization Journal.1, page 4-7.

[3] R.F. Heller et al., (1984). "How well can we predict coronary heart disease? Findings in the United Kingdom Heart Disease Prevention Project," British Medical Journal.

[4] Shashikant Ghumbre, Chetan Patil, and Ashok Ghatol(2011). "Heart Disease Diagnosis using Support Vector Machine".

[5] Mai Shouman, Tim Turner, and Rob Stocker (2012)." Applying k-Nearest Neighbor in Diagnosing Heart Disease Patients".

[6] Australian Bureau of Statistics. (July 2017-Febuary 2011). [Online]. Available: http://www.ausstats.abs.gov.au/Ausstats/subscriber.nsf/0/E8510D1C8DC1AE1CCA2576F600139288/$File/33030_2008.pdf

[7] Donna L. Hudson and Maurice E Cohen. (2003) "Neural Networks and Artificial Intelligence for Biomedical Engineering. IEEE Press Series on Biomedical Engineering." Wiley- IEEE Press.

[8] Prakash Mahindrakar, Dr. M. Hanumanthappa (2012)," Data Mining in Healthcare: A Survey of Techniques and Algorithms with Its Limitations and Challenges".

[9] Monali Dey, Siddharth Swarup Rautaray(2014)."Study and Analysis of Data mining Algorithms for Healthcare Decision Support System".

[10] M. Durairaj, V. Ranjani(2013) ."Data Mining Applications In Healthcare Sector: A Study"

[11] Hao Zhang Alexander C. Berg(2011)."SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition"

[12] V. Krishnaiah, G. Narsimha & N. Subhash Chandra, (Mar 2013) A Study on Clinical Prediction using Data Mining Techniques, International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) ISSN 2249-6831 Vol. 3, Issue 1, 239-248 TJPRC Pvt. Ltd.