K-means based SVM for Prediction Analysis

Shubhanshi Singhal[1], Pooja Sharma[2], Vishal Passricha[3]
[1]Assistant Professor, TERii, Kurukshetra, Haryana, India-136119
[2]Lecturer, Government College for Women, Karnal, India-132001
[3] Assistant Professor, National Institute of Technology, Kurukshetra, Haryana, India- 136119
(E-mail: shubhanshi17@gmail.com[1], sharmapuja42@gmail.com[2], vishal_pasricha@yahoo.com[3])

***Abstract* --** The data mining is the technique which is applied to extract useful information from the rough data. The prediction analysis methods are used to predict the future values from the current available data. This research mainly focused on the prediction analysis using the techniques of classification. In existing techniques, support vector machine (SVM) classifier shows good results for the prediction analysis. In this paper, we proposed a new model that combines both structured and unstructured classifier i.e. K-means based classifier is used to customize the data for SVM to get the better results. The performance of K-means SVM classifier-based prediction system is evaluated on UCI heart disease dataset. The proposed method is implemented in MATLAB and results are analyzed in terms of accuracy and execution time.

***Keywords*--** Data Mining, Heart Disease, K-means, Prediction analysis, SVM,

I. INTRODUCTION

There is huge amount of data being generated by various applications that need to be stored in data storage in an arranged manner. In order to analyze such huge data, powerful approaches or tools are required. Such approaches can achieve interesting knowledge for the users using decision making [1]. Thus, the method like data mining is applied on implicit, unknown and highly useful data. The process of extracting knowledge from the huge storage databases is known as data mining or Knowledge Discovery in Databases (KDD). Data mining is known as an important part of the KDD process even though they are considered to be synonyms by users. The valuable information from the huge databases is extracted with the help of similarities present amongst this data, statistics, and search strings. Data dredging, knowledge extraction, and pattern discovery are the other names used for data mining [2]. The descriptive and predictive types of data mining tasks are the two broader categories. The general properties of existing data are described with the help of data mining [3]. The predictions that are made on the basis of inference on available data are known as prediction data mining. The gathering and managing of data along with its analysis and prediction are done within data mining. Although machine learning approaches like restricted Boltzmann machines, deep neural networks, SVM are being used within several applications. Their most popular application is data mining. Lots of issues arise when the relationships amongst multiple features are analyzed by people. The identification of appropriate solutions is very difficult. machine learning approaches can be applied successfully to enhance the efficiency of the system. In this world, billions of people are suffering from heart disease. Early prediction of heart disease always plays an important role in the control and diagnosis. By early prediction, an efficient and more accurate treatment is also offered to the patient.

In this paper, both structured and unstructured model are combined in the field of healthcare to reduce the risk of heart disease. A new model is proposed in which k-means algorithm is used to customize the data and SVM works on that specified cluster to predict the results. The performance of the new model is evaluated on the UCI heart disease dataset. Through this experiment, we draw a conclusion that the accuracy of the proposed model is better than other existing methods.

II. LITERATURE REVIEW

Some of the existing models that are proposed for prediction of heart disease are given in table 1.

Table 1: Summary of some existing heart disease prediction model

Author	**Purpose**	**Techniques Used**	**Tool**	**Accuracy**
Florence et al. [4]	This system is used to predict the heart attack and also discussed various uses of various data mining algorithm for disease prediction.	Convolutional Neural Network and Decision Tree	Rapid Miner	82%

Kumari and Godara [5]	The objective is to analyze various data mining techniques on cardiovascular disease dataset	Decision Tree, Neural Networks, Support Vector Machines	Weka3.6.6	84.12%
A. Taneja [6]	Their purpose is to use various data mining techniques and an attempt to assist in the diagnosis of the heart disease	Naive Bayes, Decision tree, Neural Networks	Weka 3.6.4	89%
P. Cortez [7]	It presents the use of data mining algorithm, in classification and regression tasks.	Neural Network, SVM	R tool	86%
Velu and Kashwan [8]	The main objective is diagnosis of heart disease using Multiple Kohenen Self Organizing Maps	SVM, KSOM	Orange	99.1%
Waghulde and Patil [9]	Their focus is mainly on Genetic Neural Approach for Heart Disease Prediction	Genetic-Neural Network	Matlab	98%

III. EXISTING METHODOLOGIES

A. Support Vector Machine

SVM is a popular classifier that is used in regression, classification and general pattern recognition within data mining [10]. The initial form of SVM is a binary classifier where the output of the learned function is either positive or negative [11]. There is no need to add any prior knowledge. When there is the high dimension of input space then using Kernel methods it offers better results. SVMs does the mapping from input space to feature space to support nonlinear classification problems[12]. The kernel trick is helpful for doing this by allowing the absence of the exact formulation of mapping function which could cause the issue of the curse of dimensionality. Geometric representation is the best classification function within this approach. A separating hyperplane $f(x)$ that passes through the middle of two classes is correspondent to the linear classification function in case of a linearly separable dataset [13].

B. K-Nearest neighbor

The learning performed using analogy is the base of KNN classifiers. With the help of n-dimensional numeric attributes, the description of the training samples is done. A point within the n-dimensional space is represented by each sample. Thus, within the n-dimensional pattern space, all the training samples are stored. The pattern space for k training samples which are nearest to unknown samples is searched by KNN classifier in the case when an unknown sample is given. The Euclidean distance helps in defining the closeness of samples. Thus, amongst two given points $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_n)$, the Euclidean distance can be defined by:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad \dots(1)$$

As all the computation is delayed to that particular time duration, the speed of classification becomes less. Each attribute is assigned with equal weight by nearest neighbor classifiers which are not possible in decision tree induction and backpropagation [14]. In case there are several irrelevant attributes within the data, confusion might be generated here. In order to provide a prediction, the nearest neighbor classifiers can also be utilized such that for a given unknown sample, the real-valued prediction can be returned [15]. The average values of real-values that are associated with k-nearest neighbors are returned here by this classifier. Amongst all other machine learning algorithms, the KNN is the simplest one. On the basis of the majority votes of the neighbors, an object can be classified.

IV. PROPOSED METHODOLOGY

The prediction analysis technique is used to predict the situations according to the input dataset. The prediction analysis requires two phases. In the first phase, the k-mean clustering is applied which will cluster the similar and dissimilar type of data. In the second phase, the SVM classifier is applied which will classify the data. The k-mean clustering consists of three steps. The first step, the arithmetic mean of the whole dataset is calculated which is taken as the central point. The second step, Euclidean distance is calculated for all the points from the central point. Finally, the data will be clustered according to their similarity. The clustered data will be given as input to the SVM classifier for the classification. The data classification quality depends upon the cluster quality. In this work, the k-mean clustering algorithm will improve the cluster quality which increases classification quality. Backpropagation algorithm is used to calculate the Euclidean distance in a dynamic manner and Euclidean distance at which maximum accuracy is achieved is the final distance for the data clustering.

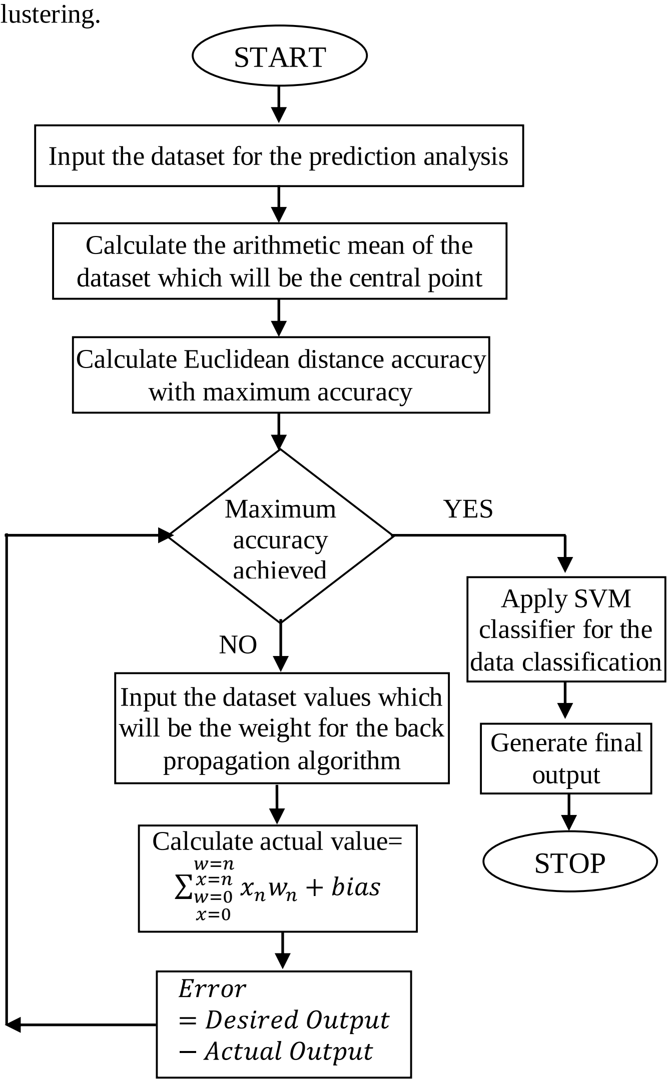

START
Input the dataset for the prediction analysis
Calculate the arithmetic mean of the dataset which will be the central point
Calculate Euclidean distance accuracy with maximum accuracy
Maximum accuracy achieved
YES
NO
Apply SVM classifier for the data classification
Generate final output
STOP
Input the dataset values which will be the weight for the back propagation algorithm
Calculate actual value= $\sum_{x=0}^{x=n}{}_{w=0}^{w=n} x_n w_n + bias$
$Error = Desired\ Output - Actual\ Output$

Fig 1: Proposed Flowchart

As illustrated in figure 1, the dataset for the classification is taken as input and central point is calculated by taking arithmetic mean of the dataset. The Euclidean distance is calculated which define data similarity. It is calculated dynamically and final iteration is that at which maximum accuracy is achieved. The formula of (actual output – desired output) is applied which will calculate error at every iteration and when the error is reduced to minimum, the maximum accuracy is achieved. When the maximum accuracy is achieved, the SVM classifier has been applied to classify the input data.

V. RESULTS AND DISCUSSION

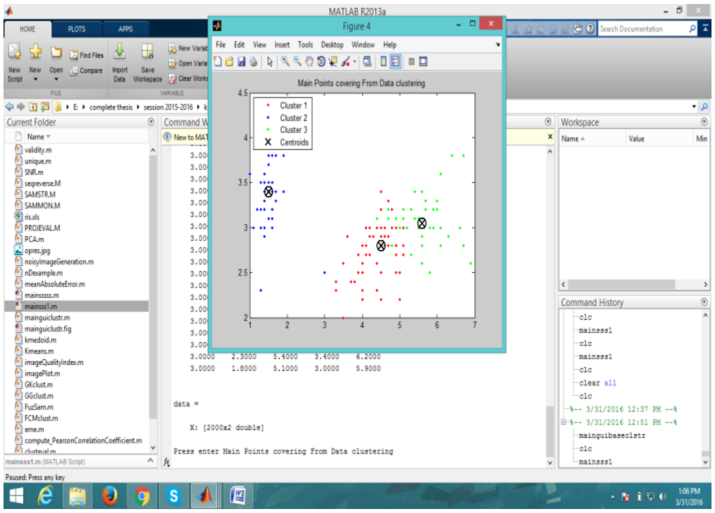
The proposed algorithm has been implemented in MATLAB-2017b by considering the dataset which is described in table 2 and the description of the attributes used for the experiment is given in table 3.

Table 2: Dataset Parameters

Parameter	Values
Dataset	Heart Disease (UCI)
Number of attributes	14
Missing Values	No
Prediction values	0 (no disease) and 1 (Heart Disease)
Parameter	Values

Table 3: Attributes used in the prediction model

Clinical Features	Description
Age	Age
Ca	Number of major vessels (0-3) colored by fluoroscopy
Chol (mg/dl)	Serum Cholesterol
Cp	Chest Pain type
Exang	Exercise-induced angina
Fbs	Fasting blood sugar
Num	Diagnosis of heart disease
Oldpeak	ST depression induced by exercise relative to rest
Restecg	Resting electrocardiographic results
Sex	Gender
Slope	The slope of the peak exercise ST segment
Thal	3=normal; 6=fixed defect; 7= reversible defect
Thalach	Maximum heart rate achieved
Trestbps (mmHg)	Resting Blood Pressure

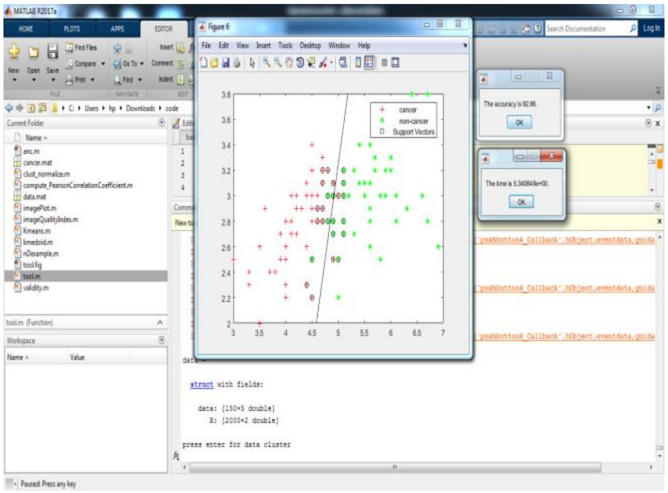


```
3.00
3.00
3.00
3.00
3.00
3.00
3.00
3.00
3.00
3.00
3.00
3.00
3.00
3.00
3.00
3.00
3.00
3.0000    2.3000    5.4000    3.4000    6.2000
3.0000    1.8000    5.1000    3.0000    5.9000


data =

    X: [2000x2 double]

Press enter Main Points covering From Data clustering
```

```
clc
mainsss1
clc
mainsss1
clc
clear all
clc
%-- 3/31/2016 12:37 PM --%
%-- 3/31/2016 12:51 PM --%
mainguibaseclstr
clc
mainsss1
```

Fig 2: Data Clustering

As shown in figure 2, the k-means algorithm is applied with the back propagation for the data clustering.



The accuracy is 92.86.

The time is 5.340848e+00.

```
'pushbutton4_Callback',hObject,eventdata,guida
'pushbutton4_Callback',hObject,eventdata,guida
'pushbutton4_Callback',hObject,eventdata,guida
```

```
data =

  struct with fields:

    data: [150×5 double]
       X: [2000×2 double]

press enter for data cluster
```

Fig 3: Data Classification

As shown in figure 3, the SVM classifier has been applied to classify the data which is the output of data clustering.

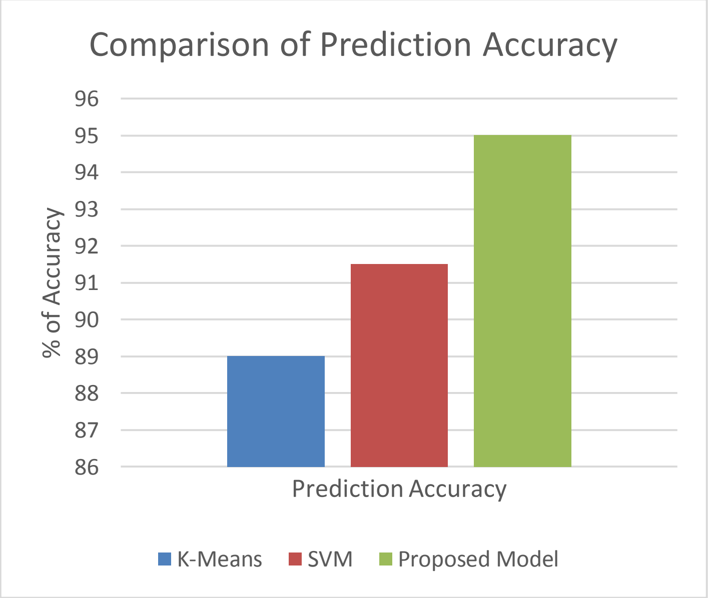


Fig 4: Accuracy Comparison

As shown in figure 4, the accuracy of the proposed and existing algorithm is compared and it is analyzed that the proposed algorithm offers high accuracy due to the clustering of uncluttered points from the dataset. In the existing algorithms, SVM classifier is directly applied to the data classification. In the proposed work, K-means and backpropagation algorithm is applied for the dynamic calculation of Euclidean distance which accurately calculates the Euclidean distance for the clustering of data.

VI. CONCLUSION

In this work, the prediction analysis technique is applied to search future possibilities from the current data. This research work is based on the prediction analysis using techniques of classification. In this existing work, only SVMs are applied for the prediction analysis. In this research, SVM based classifier's performance is improved using the K-means algorithm to provide them sorted data. The performance of both techniques is also separately analyzed in terms of accuracy and compared with the proposed system. The proposed algorithm shows high accuracy as compared to the existing algorithm.

REFERENCES

[1] K. Rajalakshmi, D. S. Dhenakaran, and N. Roobin, "Comparative Analysis of K-Means Algorithm in Disease Prediction," *International Journal of Science, Engineering and Technology Research (IJSETR)*, vol. 4, pp. 1-3, 2015.

[2] O. Oyelade, O. Oladipupo, and I. Obagbuwa, "Application of k Means Clustering algorithm for prediction of Students Academic Performance," *arXiv preprint arXiv:1002.2425*, 2010.

[3] B. SundarV, T. Devi, and N. Saravanan, "Development of a Data Clustering Algorithm for Predicting Heart," *International Journal of Computer Applications*, vol. 48, pp. 8-13, 2012/06/30 2012.

[4] S. Florence, N. B. Amma, G. Annapoorani, and K. Malathi, "Predicting the risk of heart attacks using neural network and decision tree," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, pp. 7025-7030, 2014.

[5] M. Kumari and S. Godara, "Comparative study of data mining classification methods in cardiovascular disease prediction 1," 2011.

[6] A. Taneja, "Heart disease prediction system using data mining techniques," *Oriental Journal of Computer science and technology*, vol. 6, pp. 457-466, 2013.

[7] P. Cortez, "Data mining with neural networks and support vector machines using the R/rminer tool," in *Industrial Conference on Data Mining*, 2010, pp. 572-583.

[8] C. Velu and K. Kashwan, "Visual data mining techniques for classification of diabetic patients," in *Advance Computing Conference (IACC), 2013 IEEE 3rd International*, 2013, pp. 1070-1075.

[9] N. P. Waghulde and N. P. Patil, "Genetic neural approach for heart disease prediction," *International Journal of Advanced Computer Research*, vol. 4, p. 778, 2014.

[10] V. C. Osamor, E. F. Adebiyi, J. O. Oyelade, and S. Doumbia, "Reducing the Time Requirement of k-means Algorithm," *PLoS One*, vol. 7, p. e49946, 2012.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273-297, 1995.

[12] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, pp. 293-300, 1999.

[13] A. Rauf, S. M. Sheeba, S. Khusro, and H. Javed, "Enhanced k-mean clustering algorithm to reduce number of iterations and time complexity," *Middle-East Journal of Scientific Research*, vol. 12, pp. 959-963, 2012.

[14] K. C. Agrawal and M. Nagori, "Clusters of Ayurvedic Medicines Using Improved K-means Algorithm," in *International Conf. on Advances in Computer Science and Electronics Engineering*, 2013.

[15] M. Sultana, A. Haider, and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," in *Electrical Engineering and Information Communication Technology (ICEEICT), 2016 3rd International Conference on*, 2016, pp. 1-5.

Shubhanshi Singhal received her B. Tech in Computer Science and Engineering from Utter Pradesh technical university, Lucknow, India. She did her Master in Technology in Computer Engineering with honours from Kurukshetra University, Kurukshetra. Presently she is working as Assistant Professor in Computer Science and Engineering Department, TERii, Kurukshetra. Her area of Interest is Data Deduplication and Machine Learning.

Pooja Sharma received her Bachelor of Science in Computer Application from Kurukshetra University, Kurukshetra, India in 2009. She received her master in computer application from Kurukshetra University, Kurukshetra in 2012. She is working as a lecturer in Government college for Women, Karnal since August 2013. She has been guided many undergraduates and postgraduates project. Her area of interest is Cloud Deduplication and Machine Learning.

Vishal Passricha received his B.Tech. in Computer Engineering from Kurukshetra University, India, in 2010, He received his M.Tech. with honors in Computer Engineering from YMCAUST, Faridabad India in 2012. He is currently working as an assistant professor in the Computer Engineering Department, National Institute of Technology, Kurukshetra. He is also pursuing Ph.D. from the same Department. His current research focuses on Acoustic Modeling, Speech Recognition, and Deep Learning.