ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Prediction Analysis Technique using Hybrid Clustering and SVM Classification

Tanu Minhas [1], Nancy Sehgal [2]

Research Scholar, Department of Computer Science & Engineering, Baddi University of Emerging Sciences and Technology, Himachal Pradesh, India[1]

Assistant Professor, Department of Computer Science & Engineering, Baddi University of Emerging Sciences and Technology ,Himachal Pradesh, India[2]

ABSTRACT: Data mining is a technique to accumulate, explore and analyze a large amount of complex data. Prediction Analysis is a technique which is applied to predict future outcomes and trends according to the given input datasets. In the recent time, various techniques have been applied for the prediction analysis. In this paper, BFS, CFS, k-mean, and SVM classifier based prediction analysis technique is improved to increase accuracy and execution time. In the prediction analysis based technique, Breadth first search (BFS) & Correlation based feature selection (CFS) is applied for the feature selection, the k-mean clustering algorithm is used to categorize the data and SVM classifier is applied to classify the data. The Boltzmann algorithm has been applied to the k-mean clustering algorithm to increase the accuracy of prediction analysis system. The proposed algorithm is implemented in MATLAB and it has been tested that the accuracy of clustering is increased, execution time is reduced for prediction analysis.

KEYWORDS: K-mean, SVM, Prediction, Categorization, Classification.

I. INTRODUCTION

A large amount of data which needs certain powerful data analysis tools is thus put here which is known as the data rich but information poor condition. There is an increase in the growth of data, it's gathering as well as storing it in huge databases. It is no more in the hands of humans to do it easily or without the help of analysis tools [1]. There are certain data archives created here which can be visited when the data is required. The insightful, interesting and novel patterns of data are discovered from large-scale data sets using the data mining. The knowledge discovery in databases process is a very important step in data mining. The data mining and KDD are often termed as synonyms. There are databases, data warehouses, internet, information repositories involved within the data sources. The two high-level primary goals of data mining in practice are to have a tendency for prediction and description [2].

As expressed before, prediction involves utilizing a few variables or fields as a part of the database to predict unknown or future values of different variables of interest, and description focuses on discovering human-interpretable patterns describing the data. In spite of the fact that the boundaries among prediction and description are not sharp (a portion of the predictive models can be descriptive, to the degree that they are understandable, and the other way around), the distinction is valuable for understanding the overall discovery goal. The relative importance of prediction and description for particular data-mining applications can differ considerably. The goals of prediction and description can be accomplished utilizing a variety of particular data-mining methods [3].

The data clustering is an unsupervised classification method. Its main objective is to create the group of objects or clusters in such a manner that the objects which have similar properties can be grouped together. Here the distinct objects are thus present in different groups as per their properties. Within the data mining research area, cluster analysis a very old and efficient area for study. For the purpose of knowledge discovery, this step is the starting point in this direction. The data objects are grouped into a set of disjoint classes called clusters using the clustering method. There is

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

a higher resemblance of objects which are present within the same class as compared to the two objects which belong to separate classes [4].

1.1. K-mean Clustering

The K-Means clustering utilizes a recursive system. Along these lines, its functionality is known like k-means calculation; it is characterized from the core calculation as Lloyd's calculation, especially in the Data mining community. K-means clustering is a strategy for vector quantization, initially from flag processing, that is popular for cluster investigation in data mining. K-means clustering aims to partition n observations into k clusters in which every observation has a place with the cluster with the nearest mean, serving as a prototype of the cluster [5]. This results in a partitioning of the data space into Voronoi cells. This calculation has a close relationship with the k-nearest neighbor classifier, a popular machine learning method for an arrangement that is frequently confused with k-means in light of the k in the name. One can apply the 1-nearest neighbor classifier on the cluster centers acquired by k-means to classify new data into the existing clusters [6].

1.2. SVM Classifier

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. The steps are explained below [7]:

Set up the training data: The training data of this exercise is formed by a set of labeled 2D-points that belong to one of two different classes; one of the classes consists of one point and the other of three points [8].

Set up SVM's parameters: In SVM the training examples are spread into two classes that are linearly separable. However, SVMs can be used in a wide variety of problems (e.g. problems with non-linearly separable data, an SVM using a kernel function to raise the dimensionality of the examples, etc). As a consequence of this, we have to define some parameters before training the SVM.

Regions classified by the SVM: The method which is used to classify an input sample is using a trained SVM. In this example we have used this method in order to color the space depending on the prediction done by the SVM. In other words, an image is traversed interpreting its pixels as points of the Cartesian plane. Each of the points is colored depending on the class predicted by the SVM; in green, if it is the class with label 1 and in blue if it is the class with label -1 [9].

II. LITERATURE REVIEW

Doreswamy, Umme Salma M, Medical informatics primarily deals with finding solutions for the issues identified with the diagnosis and prognosis of different deadly diseases utilizing machine learning and data mining approaches. One such disease is breast cancer, killing millions of people, particularly women. In this paper we propose a bio inspired model called BATELM which is a mix of Bat algorithm (BAT) and Extreme Learning Machines (ELM) which is first of its kind in the study of non image breast cancer data analysis. The idea of BAT and ELM which has many advantages when compared to the existing algorithms of their genre has motivated us to build a model that can predict the medical data with high accuracy and minimal error [10]. Here we make utilization of BAT to optimize the parameters of ELM so that the prediction task is completed efficiently. The fundamental aim of ELM is to predict the data with least error.

R. Karakis, M. Tez, Y. Kilic, Y. Kuru, and I. Guler, Axillary Lymph Node (ALN) status is an extremely important factor to survey metastatic breast cancer. Surgical operations which might be vital and cause some adverse effects are performed in determination ALN status [11]. The motivation behind this study is to predict ALN status by methods for selecting breast cancer patient's essential clinical and histological feature(s) that can be obtained in every healing center.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

270 breast cancer patients' data are collected from Ankara Numune Educational and Research Hospital and Ankara Oncology Educational and Research Hospital. It is concluded from LR and GA based MLP, that menopause status and lymphatic invasion are the most significant features for determining ALN status.

Marjia Sultana, Afrin Haider and Mohammad Shorif Uddin, Heart disease is considered as one of the major reasons for death all through the world. It can't be effectively predicted by the medical specialists as it is a troublesome task which demands expertise and higher knowledge for prediction. The heart disease turns into a plague all through the world. It can't be effortlessly predicted as it is a troublesome task that demands expertise and higher knowledge for prediction. Data mining extracts hidden information that assumes an important role in settling on choice [12]. This paper addresses the issue of prediction of heart disease as per input attributes on the premise of data mining strategies.

Kamaljit Kaur and Kuljit Kaur, the new system, called the Credit Based Continuous Evaluation and Grading System (CBCEGS), assesses a student on the premise of her persistent evaluation during the semester, joined with her performance at last semester examination. This multistage examination design gives a chance to students to improve their performance. In the event that a student can't perform well in tests during the semester, she can improve her performance at last semester test. In any case, it doesn't appear to be so natural [13]. In specific courses, because of their difficulty level, for example, mathematics, a student will most likely be unable to improve her knowledge at last despite hard work.

Monali Paul, Santosh K. Vishwakarma, Ashok Verma, this work presents a system, which utilizes data mining procedures keeping in mind the end goal to predict the category of the broke down soil datasets. The category, in this manner predicted will demonstrate the yielding of crops. The problem of predicting the crop yield is formalized as a classification rule, where Naive Bayes and K-Nearest Neighbor methods are utilized. In this work, classification of soil into low, medium and high categories are finished by adopting data mining procedures keeping in mind the end goal to predict the crop yield utilizing accessible dataset. This study can help the soil analysts and farmers to decide sowing in which land may result in better crop production [14]. The future work may aim to make more efficient models utilizing other data mining classification procedures.

III. RESEARCH METHODOLOGY

Prediction Analysis is a technique to predict the situations according to the input datasets. The prediction analysis required two phases, in the first phase the k-mean clustering is applied which will cluster the similar and dissimilar type of data. The SVM classifier is applied which will classify the data. The k-mean clustering consists of three steps, in the first step the arithmetic mean of the whole dataset is calculated which will be the central point. In the second step, Euclidian distance is calculated from the central point. In the last step, the data will be clustered according to their similarity. The clustered data will be given as input the SVM classifier for the classification. The data classification quality depends upon the cluster quality. In this work, the k-mean clustering algorithm will be improved to increase cluster quality which increase classification quality. The Boltzmann algorithm is applied with the k-mean clustering algorithm which increase cluster quality. The Boltzmann algorithm will calculate the Euclidian distance in the dynamic manner and Euclidian distance at which maximum accuracy is achieved is the final distance for the data clustering.

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

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

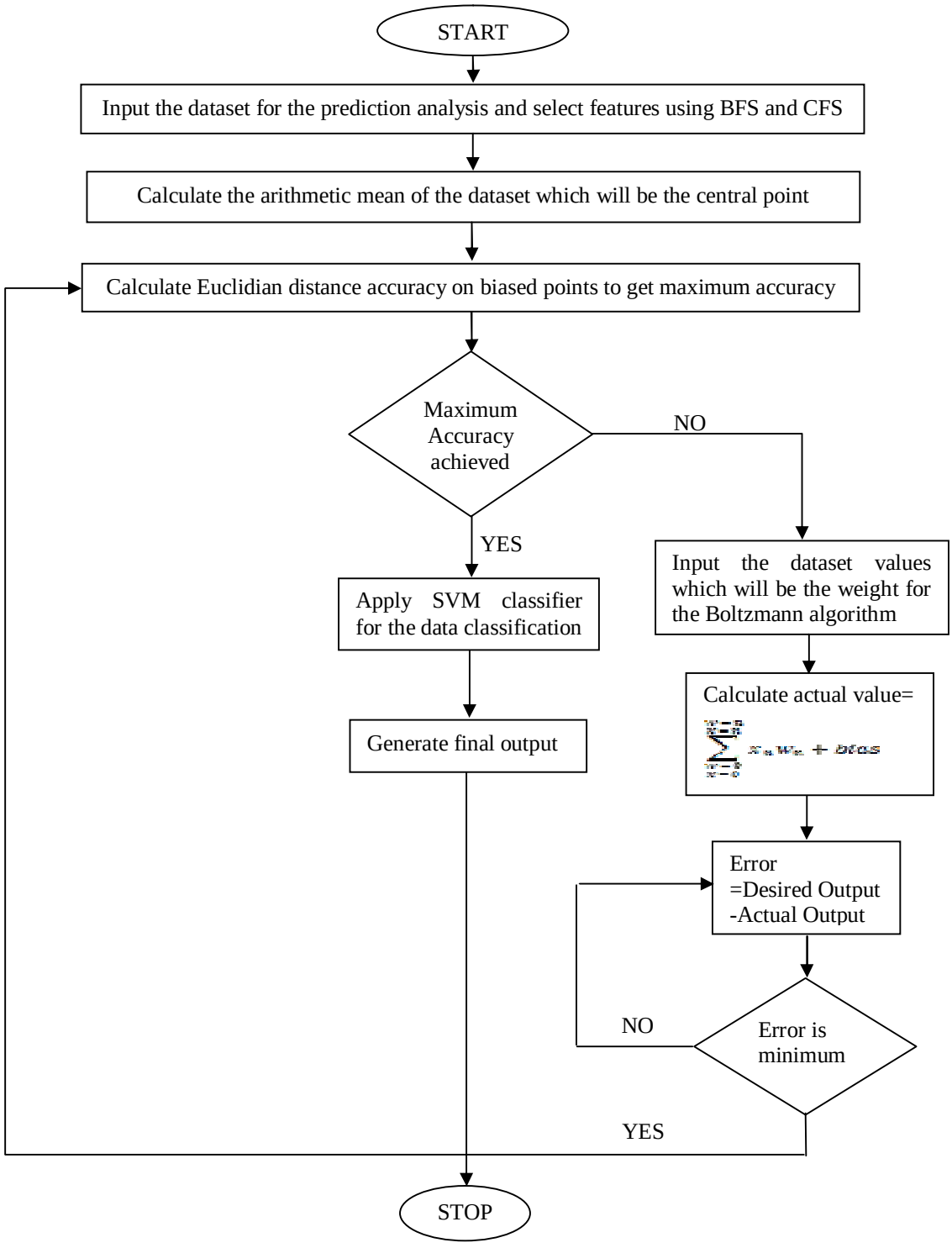

Fig: 1 Flowchart of Proposed work

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

IV. PROPOSED ALGORITHM

INPUT: Dataset

OUTPUT: Clustered Data

Start ( )

1. Read dataset and dataset has number of rows “r” and number of coloums “m”
2. Select feature from the dataset using BFO and CFO
3. For (i=0 ;i=r; i++) /// selection of medoid point
    1. For (j=0; j=m; j++)
    2. Select k=data (i, j);
    End
4. Calculation of Euclidian distance()
    1. For (i=0;i=r;i++)
    2. For (j=9;j=m;j++)
    3. A(i)=data(i);
    4. B(i)=data(j);
    5. Distance =squt[(A(i+1)-A(i)^2) –(B(j+1)-B(j)^2);
    End
5. Normalization ()
6. Weight=0,bais=0, input=0
7. R=max(x)
8. While the whole data get classified into two classes in the for loop do
9. For i=1 to CS(n) do
10. If $Y_i(<W_i X_i> +bias)< 0$ then
11. $W_{k+1}=W_k+Y_iX_i$
12. K=k+1;
13. End if
14. End while
15. Return Clustered data K, The k is the number of classes and x is the data in the classes
16. Repeat step 3 to 4 until all points get clustered.

V. EXPERIMENTAL RESULTS

The proposed algorithm has been implemented in MATLAB by considering the dataset which is described in the table1.

PARAMETER		No of instances	Attributes	Missing values	Area	Association rules
VALUES	DATASET1	569	32	NO	Life	Classification
	DATASET2	286	9	YES	Life	Classification

Table 1: Dataset Parameters

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

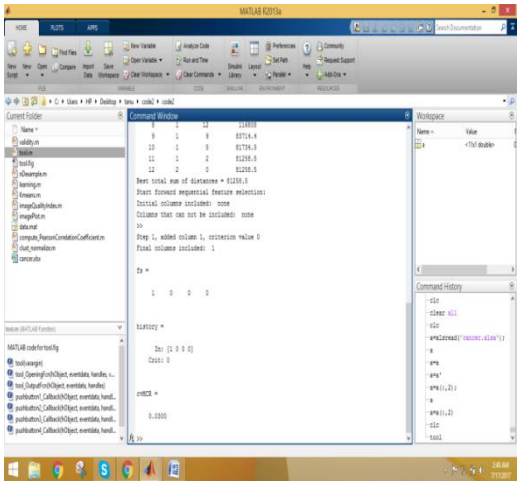

Fig 2: Apply BFS and CFS

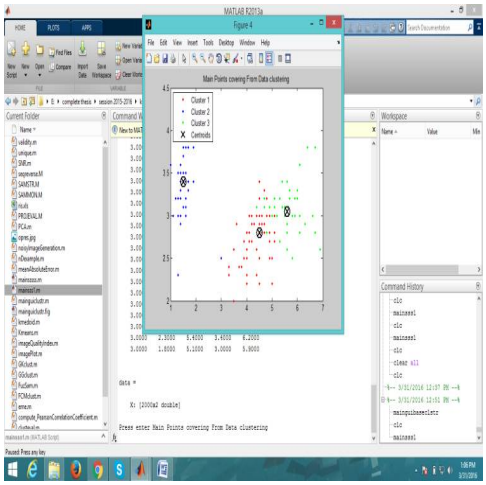

Fig 3: Data Clustering

As shown in Fig 2, the data set is taken as input from UCI repository and on that loaded data to select most relevant features BFS and CFS algorithm are applied to search and select feature respectively.

As shown in figure 3, the k-mean algorithm is applied with the Boltzmann algorithm for the data clustering.

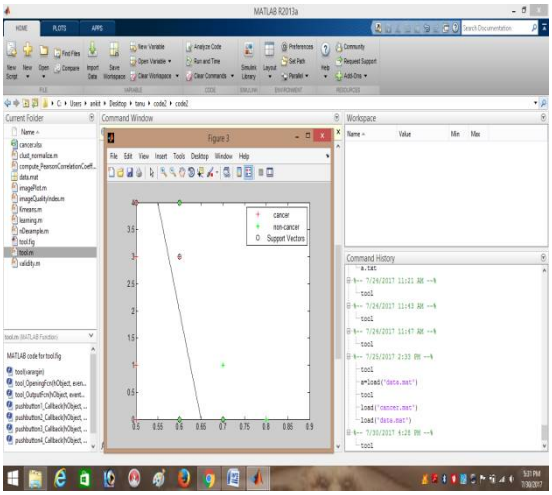

Fig 4: Data Classification

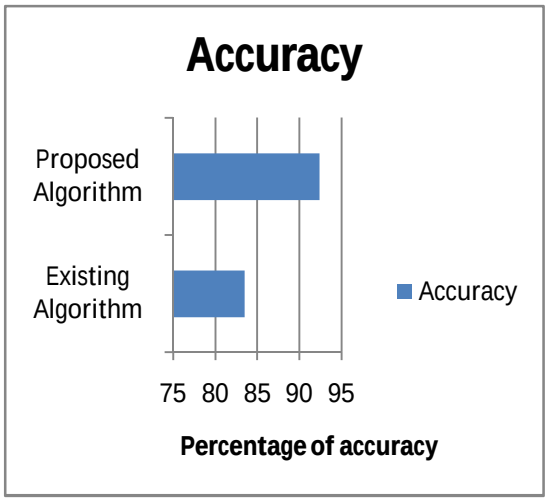

Fig 5: Accuracy Comparison

As shown in Fig: 4, the SVM classifier has been applied which will classify the data which is output of data clustering.

As shown in Fig:5, the accuracy of proposed and existing algorithm is compared and it has been analysed that proposed algorithm has high accuracy as compared to the existing one due to the use hybrid model for clustering the dataset.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

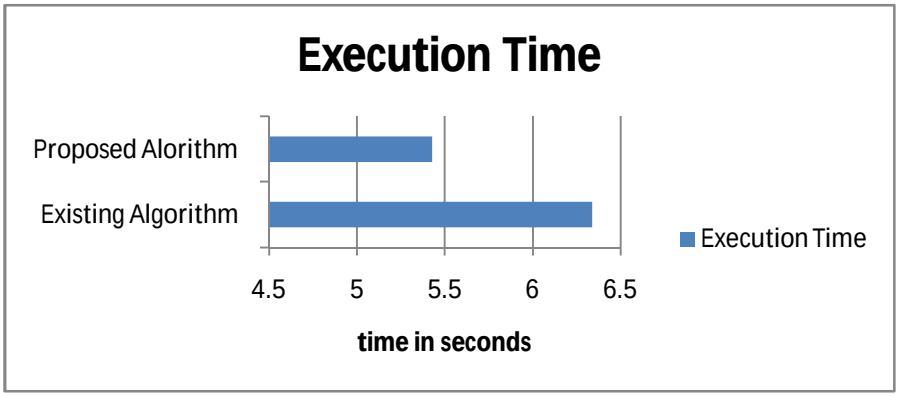

Fig 6: Execution time

As illustrated in Fig: 6, the execution time of proposed and existing algorithm has been compared and due to used of Boltzmann algorithm execution time is reduced in the proposed work.

V. CONCLUSION

In this paper, it has been concluded that prediction analysis is an efficient technique for the complex data analysis. In this work, the dataset for the prediction analysis is taken from the UCI repository onto which Breadth first search (BFS) and Correlation based feature selection (CFS) is applied for the feature selection. Then Boltzmann algorithm is applied with the k-mean clustering algorithm to increase accuracy of data clustering. In the last phase SVM classifier is applied which will classify clustered output. The performance of proposed algorithm is tested in MATLAB and it has been analysed that accuracy is increased upto 20 percent and execution time is reduced upto 10 percent.

REFERENCES

[1] Rupali, R.Patil, "Heart disease prediction system using Naive Bayes and Jelinek-mercer smothing," 2014, International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 5
[2] Shamsher Bahadur Patel, Pramod Kumar Yadav and Dr. D. P. Shukla, "Predict the diagnosis of heart disease patients using classification mining Techniques," 2013, IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS), vol. 4, no. 2, pp. 61-64
[3] Jyoti Soni, Ujma Ansari and Dipesh Sharma, "Prediction data mining for medical diagnosis: An overview of heart disease prediction," 2011, International Journal of Computer Applications (0975-8887), vol. 17
[4] John G. Cleary and Leonard E. Trigg," K: An Instance-based learner using an entropic distance measure," 1995, Proc. 12th International Conference on Machine Learning, pp. 108-114
[5] S. Vijayarani and M. Muthulkshmi, "Comparative analysis of Bayes and Lazy classification algorithms," 2013, International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, no. 8
[6] R. Vijaya Kumar Reddy, K. Prudvi Raju, M. Jogendra Kumar, CH. Sujatha, P. Ravi Prakash, "Prediction ofheart disease using decision tree approach," 2016, International Journal of Advanced Research in Computer Science and Engineering, vol. 6, no. 3
[7] Promad Kumar Yadav, K. L. Jaiswal, Shamsher Bahadur Patel, D. P. Shukla, "Intelligent heart disease prediction model using classification algorithms," 2013, UCSMC, vol. 3, no. 08, pp. 102-107
[8] Gaurav Taneja and Ashwini Sethi, "Comparison of classifiers in data mining," 2014, International Journal of Computer Science and Mobile Computing, vol. 3, pp. 102-115
[9] Sheweta Kharya, "Using data mining techniques for diagnosis of cancer disease," 2012, UCSEIT, vol. 2, no. 2
[10] Doreswamy, Umme Salma M," BAT-ELM: A Bio Inspired Model for Prediction of Breast Cancer Data", 2015, IEEE
[11] R. Karakis, M. Tez, Y. Kilic, Y. Kuru, and I. Guler," A genetic algorithm model based on artificial neural network for prediction of the axillary lymph node status in breast cancer," 2013, Engineering Applications of Artificial Intelligence, vol. 26, no. 3, pp. 945–950
[12] Marjia Sultana, Afrin Haider and Mohammad Shorif Uddin," Analysis of Data Mining Techniques for Heart Disease Prediction", 2016, IEEE
[13] Kamaljit Kaur and Kuljit Kaur," Analyzing the Effect of Difficulty Level of a Course on Students Performance Prediction using Data Mining", 2015 1st International Conference on Next Generation Computing Technologies (NGCT)
[14] Monali Paul, Santosh K. Vishwakarma, Ashok Verma, " Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining Approach", 2015, IEEE