

Back Propagation Method for Heart Disease Prediction

Surendra Singh¹, Dr. Vishal Shrivastava²

¹M.Tech. Scholar, Department of Computer Science & Engineering, ACEIT, Jaipur, Rajasthan, India

²Prof., Department of Computer Science & Engineering, ACEIT, Jaipur, Rajasthan, India

Corresponding Author; Surendra Singh

Abstract : In order to scrutinize difficult information, data mining method is utilized while prediction analysis is the method used for the prediction of information on the basis of input dataset. In the last few years, a number of methods have been implemented for prediction scrutiny. In this study, the k-means clustering algorithm and SVM (support vector machine) classifier relied prediction analysis method is utilized for clustering and classification of the key information in order to predict heart diseases. The back propagation algorithm is implemented in association with k-means clustering algorithm for the clustering of information to enhance the accurateness of prediction analysis. The performance of applied algorithm is experienced in the heart disease data sample which is gathered from UCI depository. This database carries seventy six different qualities. Though, a subgroup of fourteen qualities is necessary for every test. Specially, Cleveland database is utilized by machine learning investigators always. The projected research is compared with the accessible methods (using arithmetic mean) as well by means of correctness, error discovery rate and implementation time.

Keywords: SVM, Back propagation, Heart disease Prediction

Date of Submission: 07-07-2019

Date of acceptance: 25-07-2019

I. Introduction

The procedure used for the extraction of useful data from the unprocessed data is known as data mining. This arbitrarily retrieved information can be set in the controlled form designate to use as significant information in different functions. This procedure of information withdrawal is identified as misnomer as well. Presently, huge quantity of information is present in almost each area. The evaluation of complete information is an extremely complex task as it requires large time. This accessible information is in unprocessed form which cannot be utilized. Therefore an appropriate data mining procedure is essential for the extraction of useful information [1]. A number of economical, easier and more efficient elucidations are present for the scrutiny of the uncomplicated information. The major aim of utilizing data mining is the discovery of significant data obtainable in indistinct way. The data mining technologies can sweep and recognize earlier concealed prototypes by using databases. Information access input faults are symbolized by the prototype detection issues for example network scheme and fake credit card transactions discovery. Thus, the outcome has to be offered in such a way that human beings can recognize it easily. The analytical data can be retrieved from different applications through proficient data mining technologies [2]. In these days, the major sources of information are satellite images, company transactions, text-reports, army acumen and scientific information. All these data should be managed carefully. The information recovery procedure does not provide any suitable outcomes for decision-making. Therefore, the invention of novel techniques is necessary for handling huge quantity of information. This provides aid in the creation of good decisions. The discovery of novel prototypes and significant information is necessary for unprocessed information which reviews the whole retrieved information. The data mining procedure provides huge support in various applications. Different businesses like communication, monetary, retail and advertising associations are using this method for minimizing their job stress [3]. The dealers use data mining scheme for the growth and endorsement of their goods. With the help of data mining approach, they can make a testimony of all clients according to their acquisitions and evaluations. The data mining procedure plays an important role when enumeration of all applications becomes impossible. Several main applications of this method are cluster scrutiny, picture processing, market investigation, information scrutiny and prototype detection. The consumers are classified into clusters and prototypes in clustering method which describe the interest of a consumer with the help of marketers [4]. This technique is used in biology as well because it develops the plant and animal classifications and also classifies genes of analogous functionalities. This method is utilized for the identification of analogous houses and lands regions in geology domain. The data mining approach uses Supervised and unsupervised learning methods for the prediction of heart infections [5]. A training suite is used in supervised learning in order to learn the parameters of the mock-up whereas in case of unsupervised learning, no training suite is used such as k-means clustering. The major objectives of the data mining approach are classification and forecasting. The classification schemes

are used for the classification of information process rate and unordered values whereas prediction or forecasting schemes are used for the prediction of incessant value. Decision trees and Neural Networks are some illustrations of classification schemes while regression, association rules and clustering are the illustrations of prediction model. Decision trees, neural networks, and Naive Bayes Classifier are some classification schemes which can be utilized in the data mining for the heart disease forecasting. The decision tree algorithm is considered the most influential classification method of data mining [6]. In this technique whole models are constructed build in the shape of tree arrangement. Data suites are divided into little suites and provide aid in the formulation of a linked decision tree. In the Neural network huge amount of rudiments are arranged in dissimilar quantity of layers which are interlinked tone another. The adaptive non-linear information processing algorithms are executed with the help of this scheme which help in the integration of all multi-processing elements. These networks are described according to the self-association and usual variation [7]. The Naive Bayes classifier is an uncomplicated probabilistic classifier that relies on Bayes theorem. This algorithm is recognized as the autonomous feature model as well. The Naive Bayes classifier model is relied on the supposition that attributes of an exacting class does not relate to the current attribute of any other class. Naive Bayes classifiers are prepared to provide service in supervised learning scheme.

II. Literature Review

BayuAdhi Tama, et.al (2016) proposed a research work about persistent disease recognized as diabetes. This disease was found very usually and caused key causalities. According to the International Diabetes Federation (IDF), approximately 285 million people all across the world were suffering from diabetes disease [8]. This number could rise in the coming years as there is no suitable technique up to now for the minimization and prevention of this disease. Type 2 diabetes (TTD) was considered the most universal kind of diabetes. The main concern was the discovery of TTD because it was not simple to forecast all the effects. Thus, data mining approach was utilized since it provided the best outcomes and provided aid in the information detection from available data. The support vector machine (SVM) algorithm was used in the data mining procedure which acquired all the information through the extraction of patients' data from earlier reports. The timely discovery of TTD helped in effectual assessment.

Yu-Xuan Wang, et.al, (2017) examined different applications which provided importance of the data mining and machine learning approaches in various domains [9]. Investigation on the organization plans of various elements of the scheme was projected because most of the work was carried out on the features of the scheme which varied timely. For the designing of novel scheme, the performance of the method was evaluated through projected technique. In this study, a novel technique was proposed for the architecture of operating system which utilized data mining and machine learning approaches. After obtaining a respond from a data miner, whole information gathered from the arrangement was scrutinized. On the basis of implemented tests, it was identified that proposed technique provided effectual outcomes.

ZhiqiangGe, et.al, (2017) offered an assessment on accessible data mining and analytical applications which were utilized in business for different functionalities. Eight unsupervised and ten supervised learning algorithms were used for the analysis objective in data mining and analytical applications [10]. An application type was presented in the study for the semi-supervised learning algorithms. In the business procedure, both unsupervised and supervised machine learning techniques were utilized extensively for around 90%-95% of whole applications. The semi-supervised machine learning approach was commenced currently. Thus, it was identified that data mining approach and forecasters played a crucial role in the scheming of novel machine learning methods for the business functionalities.

P. Suresh Kumar, et.al (2017) presented a scheme which removed numerous concerns being experienced in clustering and classification techniques. This scheme was implemented in obtainable data mining structures. This technique was utilized for the diagnosis of different kinds of diabetes disease. The gathered information was used for the calculation of health level of each patient. This disease caused various causalities because of which a number of investigations were performed in this field [11]. The information collected from approximately 650 patients was utilized for the research work. This information helped in the identification of the effects of this disease. This grouped data suite was utilized in the form of input in the classification system which was auxiliary used for the classification of information into definite classes such as risk levels of diabetes disease was categorized into serene reasonable and rigorous. The performance scrutiny of various algorithms was performed for the diagnosis of diabetes. The performance of every classification algorithm was measured according to the attained outcomes.

Han Wu, et.al (2018) presented a new scheme on the basis of data mining methods in order to predict type 2 diabetes mellitus (T2DM). The major aim of this study was to advance the correctness of the forecasting scheme and the development of an adaptive data suite system. The projected scheme was made up of two branches according to the sequence of preprocessing measures [12]. These two branches were identified as enhanced K-means algorithm and the logistic regression algorithm. The Pima Indians Diabetes Dataset and the

Waikato Environment was utilized for Knowledge Analysis toolkit in order to compare outcomes with other techniques. On the basis of conducted researches, it was analyzed that presented scheme showed enhanced correctness in comparison with other techniques and provided the adequate data suite feature as well. For evaluating the performance of proposed scheme, this scheme was implemented in other diabetes data suite too, where these schemes demonstrated high-quality performance.

Jahin Majumdar, et.al, (2016) proposed a research on data mining and machine learning approaches. These were the most popular investigative fields in computer discipline [13]. The SFS and SBS schemes were the most favorable schemes and favored when implemented in association with forward selection method. The presented heuristic approach utilized support vector machine methods because it served correctness and computational purposes. The correctness rate of support vector machine algorithm was calculated with the help of data suite. In data mining, a number of investigations were conducted to improve the data classification and pattern recognition on the basis of feature selection. The tested results depicted that comparison amid the accessible methods was performed for the selection of most appropriate method. The presented technique removed the hypothetical restrictions of accessible algorithms.

III. Research Methodology

This investigative study was relied on the forecast scrutiny of heart infections. The prediction or forecast scrutiny is the method in which prospect potentials can be forecasted on the basis of present data suite. In this investigative study, the support vector machine (SVM) algorithm is implemented formerly for the prediction scrutiny. The SVM algorithm is considered one of the easiest algorithms amid all the machine learning approaches. The decision tree is identified as non-parametric supervised learning algorithm as no suppositions are made on the fundamental information allocation. In this approach, the patterns are classified according to the adjacent training patterns existing inside the characteristic gap. The characteristic vectors are amassed together with the labels of training images during the training procedure. The unlabelled query end is doled out during the categorization procedure towards the label of its k-nearest neighbors. The object is described with the help of bulk share vote on the base of its associates' labels. The entity is classified fundamentally because the class of the entity closest to it in the event when $k=1$. k is identified as odd integer in a situation, when only two classes are present. For the duration of the performance of multiclass classification, there can be tie in case when k is identified as an odd whole digit.

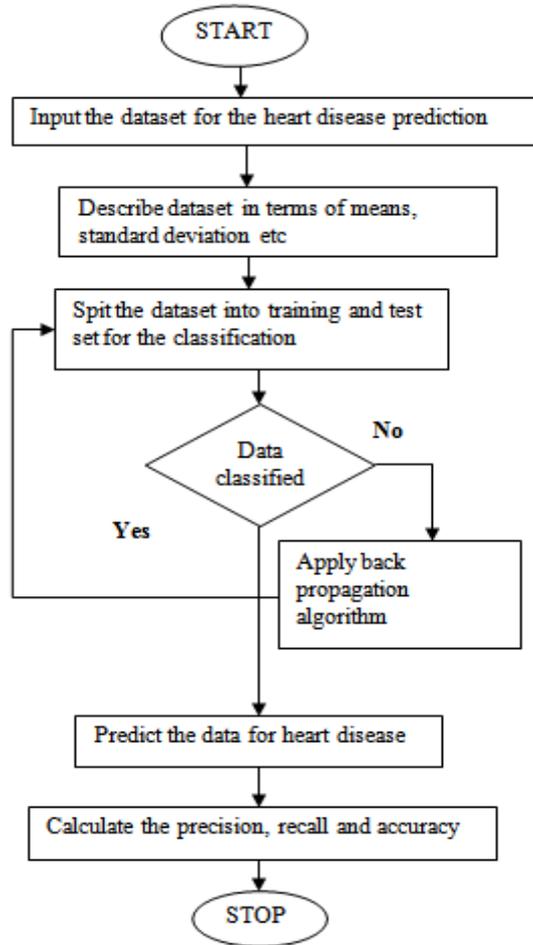


Fig 1: Proposed Methodology

IV. PERFORMANCE EVALUATION

The presented scheme is executed in Python and the outcomes are examined with the help of comparisons performed amid presented and obtainable schemes by means of correctness and implementation time. The dataset is collected from the UCI repository. The details of the dataset is described in table 1

Table 1 : Dataset Description

Data Characteristics:	Set	Multivariate	Number Instances:	of	303	Area:	Life
Attribute Characteristics:		Categorical, Integer, Real	Number Attributes:	of	75	Date Donated	1988-07-01
Associated Tasks:		Classification	Missing Values?		Yes	Number of Web Hits:	938288

1. Accuracy: Accuracy or correctness is described as the total number of points properly classified divided by total number of points multiplied by 100

$$\text{Accuracy} = \frac{\text{Number of points correctly classified}}{\text{Total Number of points}} * 100$$

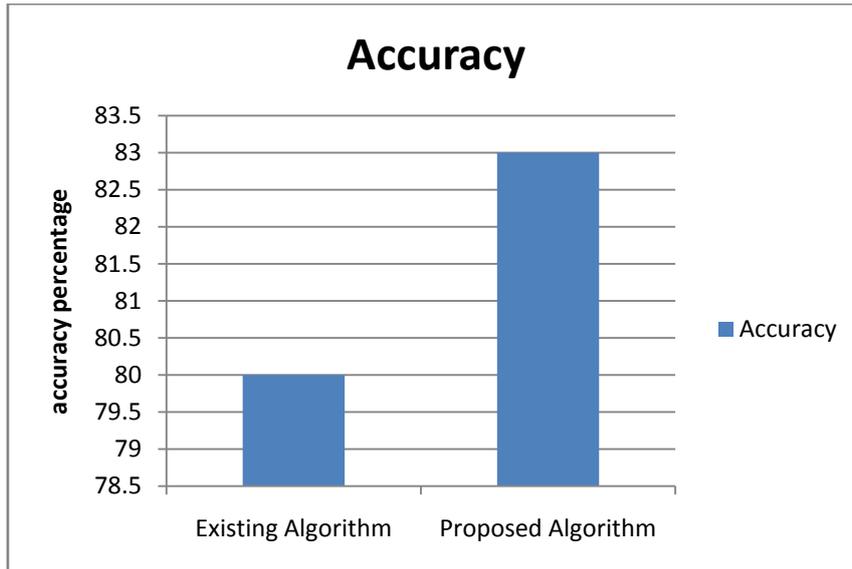


Fig 2: Accuracy Comparison

The correctness comparison of accessible and presented algorithm is demonstrated in the figure 2. The correctness of presented algorithm is superior to accessible algorithm.

2. Execution Time: Execution time or implementation is described as dissimilarity of end time when algorithm discontinues performing and starts time when algorithm starts performing as depicted in equation 2.

Execution time = End time of algorithm- start of the algorithm

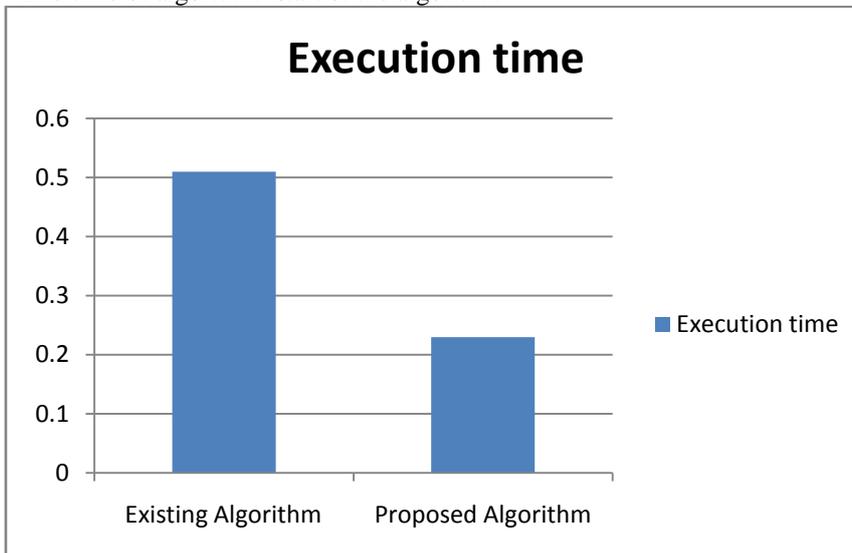


Fig 3: Execution time

The execution time of presented and accessible algorithm is demonstrated in the figure 3. The execution time of presented algorithm is smaller than the accessible algorithm.

3. CAP Analysis: -A canonical scrutiny is performed on the major coordinates for any semblance matrix, involving a variation experiment. The data structure is considered by CAP. Thus, it is more likely to divide dissimilar levels if there is no sturdy dissimilarity and is good to demonstrate the communication amid features.

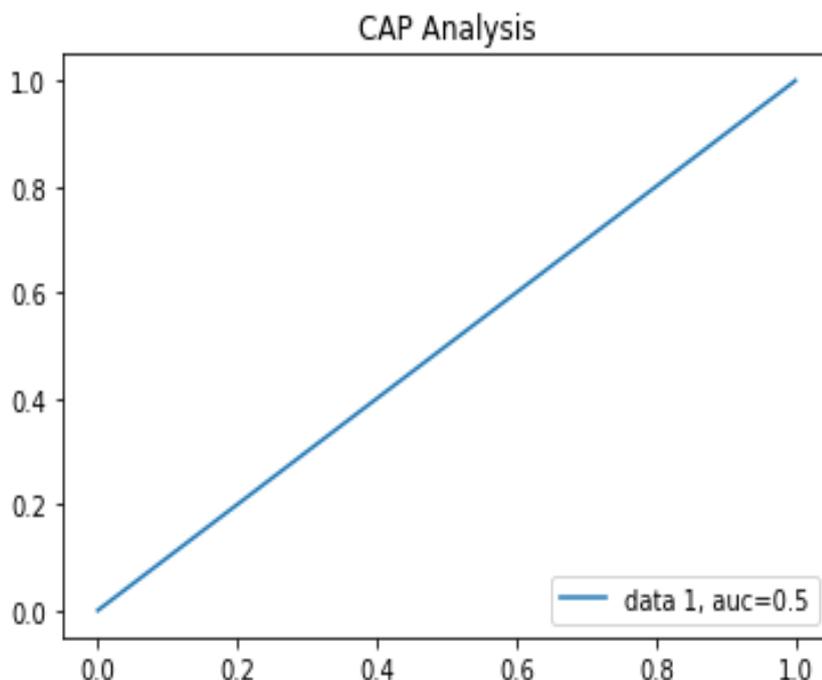


Fig 3: CAP Analysis

The CAP scrutiny is demonstrated in this figure 3. The training data suite is applied in the form of input on the x-axis of this curve and on the y-axis the test data is applied in the form of input. The blue line depicts the CAP curve which symbolizes the correctness of the classification model.

V. Conclusion

The pertinent data is extracted from coarse data suite with the help of data mining method. The alike and unlike information is clustered after computing a resemblance amid input data suite. The support vector machine (SVM) classifier is utilized for the classification of both alike and unlike data kinds where centralized point is measured through the calculation of an arithmetic mean of the data suite. The centralized point computed the Euclidian distance is utilized for the computation of resemblance amid dissimilar information points. The clustered data is classified with the help of SVM classification model on the basis of input data suite kind. In this investigative study, back propagation algorithm is implemented along with SVM classification model for increasing the correctness of forecasting. The presented algorithm gives good performance by means of correctness and implementation time. As future work, the presented technique will be enhanced auxiliary for the scheming of hybrid classification model in order to predict heart disease.

References

- [1]. Yanhui Sun, Liying Fang and Pu Wang, Improved k-means clustering based on Efros distance for longitudinal data, 2016 Chinese Control and Decision Conference (CCDC), Vol. 11, issue 3, pp. 12-23, 2016.
- [2]. Shunye Wang, Improved K-means clustering algorithm based on the optimized initial centroids, 2013 3rd International Conference on Computer Science and Network Technology (ICCSNT), Vol. 11, issue 3, pp. 12-23, 2013.
- [3]. PhattharatSongthung and KunwadeeSripanidkulchai, Improving Type 2 Diabetes Mellitus Risk Prediction Using Classification, 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), Vol. 11, issue 3, pp. 12-23, 2016.
- [4]. Jiawei Han, MichelineKamber, "Data Mining: Concepts and Techniques", vol. 3, pp. 1-31, 2000.
- [5]. Ms. Tejaswini U. Mane, "Smart heart disease prediction system using Improved K-Means and ID3 on Big Data", 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), vol. 8, issue 11, pp. 123-148, 2017.
- [6]. SellappanPalaniappan, RafiahAwang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", vol. 5, issue 1, pp. 13-28, 2008.
- [7]. KanikaPahwa, Ravinder Kumar, "Prediction of Heart Disease Using Hybrid Technique For Selecting Features", 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), vol. 4, issue 5, pp. 23-48, 2017.
- [8]. BayuAdhi Tama,1 Afriyan Firdaus,2 Rodiyatul FS, "Detection of Type 2 Diabetes Mellitus with Data Mining Approach Using Support Vector Machine", Vol. 11, issue 3, pp. 12-23, 2008.
- [9]. Yu-Xuan Wang, QiHui Sun, Ting-Ying Chien, Po-Chun Huang, "Using Data Mining and Machine Learning Techniques for System Design Space Exploration and Automatized Optimization", Proceedings of the 2017 IEEE International Conference on Applied System Innovation, vol. 15, pp. 1079-1082, 2017.

- [10]. ZhiqiangGe, Zhihuan Song, Steven X. Ding, Biao Huang, "Data Mining and Analytics in the Process Industry: The Role of Machine Learning", 2017 IEEE. Translations and content mining are permitted for academic research only, vol. 5, pp. 20590-20616, 2017.
- [11]. P. Suresh Kumar and V. Umatejaswi, "Diagnosing Diabetes using Data Mining Techniques", International Journal of Scientific and Research Publications, Volume 7, Issue 6, June 2017.
- [12]. Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang, "Type 2 diabetes mellitus prediction model based on data mining", ScienceDirect, Vol. 11, issue 3, pp. 12-23, 2018.
- [13]. JahinMajumdar, Anwasha Mal, Shruti Gupta, "Heuristic Model to Improve Feature Selection Based on Machine Learning in Data Mining", 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), vol. 3, pp. 73-77, 2016.

Surendra Singh" Back Propagation Method for Heart Disease Prediction" International Journal of Engineering Science Invention (IJESI), Vol. 08, No. 07, 2019, PP 26-32