IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 08, Issue 12 (December. 2018), ||V (II) || PP 67-73

A Novel Machine Learning Approach to Predictions in Heart Disease Using Iaca

Susmitha K[1], B. Senthil Kumar[2]
M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India[1]
Assistant Professor, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India[2]
Corresponding Author: Susmitha K

Abstract: Data mining is applied in almost all type of applications like ecommerce, many business, education and health care, etc. The most famous area of data mining is handling health care datasets. From the numerous health attributes, disease prediction and its risk analysis are performed by effective data mining techniques such as clustering and classification. Detection of heart diseases from patient electronic health records is very dynamic in nature and achieving that is a very promising area of research in now days. From the numerous health data, the proposed system handles popular disease dataset such as heart diseases. In the proposed system IACA algorithm has been proposed to address the active learning problem, and this created with the aim at detecting the object label from a large amount of data. The existing system suffers from clustering problem in large scale dataset, to handle such issues, the system concentrates on three main portions for accurate clustering. One is Effective pre-processing, feature selection and clustering. The pre-processing stage eliminates the inconsistent and redundant dataset. The second stage is the feature selection process, which performed using feature sequence selection and effective clustering using IACA (Incremental advanced Clustering Algorithm). The system developed with the intension of high clustering accuracy and less time interval.

Date of Submission: 29-11-2018 Date of acceptance: 13-12-2018

I. INTRODUCTION

In this paper is to develop a prototype for Heart Disease Prediction System (HDPS) using three data mining modeling techniques namely, Decision Trees, Naive Bayes and Neural Network. Heart Disease prediction system (HDPS) is to provide knowledge and give information is associated with heart disease to historical heart disease database. In this paper can give information or complex queries for diagnosing heart disease and thus assist healthcare practitioners to make intelligent healthcare decisions which traditional decision support systems cannot**.** It is to elaborate visualization and ease of interpretation, it displays the results both in tabular and graphical forms. Heart Disease prediction diagnosis is regarded as an important yet difficult task that needs to be executed accurately and efficiently. Unfortunately all doctors do not possess knowledge in a particular field in every sub specialty and moreover there is a shortage of resource persons at certain places. Therefore, an automatic medical diagnosis system would almost certainly be exceeding beneficial by bringing all of them together. A wide variety of areas containing as part of the whole being considered marketing, customer relationship management, engineering, medicine, crime detailed examination of the elements, expert prediction, Web mining, and mobile computing, besides others utilize Data mining. It is possible to gather knowledge and solution concerning a disease from the patient specific stored measurements as far as medical data is concerned. Therefore, data mining has developed into an essential domain in healthcare. Data mining can deliver the action of assessing someone of which courses of action prove effective by note the similarity and evaluating causes, symptoms, and courses of treatments. Functioning on heart disease patient's databases is one kind of distinct from a fictional application. The detection of a disease from various factors or symptoms is a many layers problem and might lead to false thing that is accepted as true frequently associated with unpredictable effects. Therefore it especially without apparent cause reasonable to try utilizing the knowledge and experience of several specialists composed in databases towards assisting the diagnosis process.

II. PROBLEM DEFINITION

In the real time concept of heart disease prediction is very important in the medical field. Every year number of peoples are affected the heart disease problem. This problem holds simple disease to hard disease. The simple disease problem is makes the hard problem. The data mining techniques is one of the growing fields in this beginning problem coming time. Then it will moved into the next level predict all diseases in the heart

and predict the hard heart problems, but all requirements of the prediction of heart disease problem is fulfill the data mining area, so it must of the prediction process.

In [1] **S. Sharmila** has presents the concept is, "Analysis of heart disease prediction using data mining techniques". In this concept use some mathematical techniques and predicts the heart disease. The technique is almost uses the term classification. Use decision trees, linear programming, neural network and the statistics based predicts the heart disease it is that the concept of this paper. In data mining prediction that discovers the relationship between the independent variables and the relationship among dependent and the independent variables.

The concept [2], **Aqueel Ahmed at el.** has presents and shares the research experiences are organized by namely, "data mining techniques to find out heart diseases". The author is makes different reasons of these heart diseases. The field of medicals considers one of the major issues is this heart disease. Heart diseases are caused by **morbidity and mortality** in the modern society. The huge amount of the medical data availability is leads to the needs for the powerful data analyzing tools to extracts the useful knowledge. The data mining techniques are implemented in this term and to get the better application. The data mining main process they are used in this area that is the KDD process. The process of knowledge discovery and the data mining can have founds the numerous application in the business and scientific domains. The data mining techniques and decision tree and SVM is most effective terms for the heart disease. The term data mining could help in the identification or the prediction of high or low risk heart disease.

In [3] **Ramin Assari et al.** has presents the "Heart Disease Diagnosis Using Data Mining Techniques", which means it mostly focuses on the preventable and controllable disease called the heart disease. According to the World Health Organization (WHO) at the beginning stage and timely diagnosis the heart disease and it considers the remarkable role in prevents the disease progress and reducing related treatment cost. The growth of data mining is solves this problems.

In [4] **Mudasir M Kirmani** has presented the heart disease related paper namely, "Cardiovascular disease Prediction using Data mining Techniques: A Review". In this concept first discusses the basic reason on the heart disease then how to detect from that disease. The heart disease mostly coming reasons are the behavioral and food habits such as tobacco intake, unhealthy diet and obesity, physical inactivity, ageing and addiction to drugs and alcohols and that factors such as the hypertension, diabetes, hyperlipidemia, stress and other ailments, these are give high risk heart disease problem called, cardiovascular diseases. If the data mining technique are implemented in this area is provides the better and reliable prediction and diagnosis of the heart diseases. The data mining technique for heart disease is gives the methodologies are, decision tree and its variants, naïve bayes, neural networks, support vector machines (SVM), Fuzzy rules, genetic algorithms and the Ant Colony Optimization.

In [5] **Aditya Methaila et al.** has presents the paper "Early Heart Disease Prediction using Data Mining techniques". Then the classification modeling techniques are mostly following in this concept for detecting the heart disease. It follows basic techniques and apriori algorithm and the MAFIA algorithm. It collects the medical profiles and fields then can predict the heart disease.

III. PROPOSED SYSTEM

The chapter discusses about the proposed methodology and the steps involved in that. The system proposes a new iterative approach, which concentrates on the effective Feature sequence selection using Genetic Algorithm and effective clustering using Incremental advanced Clustering Algorithm (IACA).This chapter discuss about the algorithms and methodologies.

- The system implements a new Genetic based Feature sequence selection algorithm for effectively reduce the prediction delay. The system introduces a new Incremental advanced Clustering Algorithm to achieve the highest accuracy.

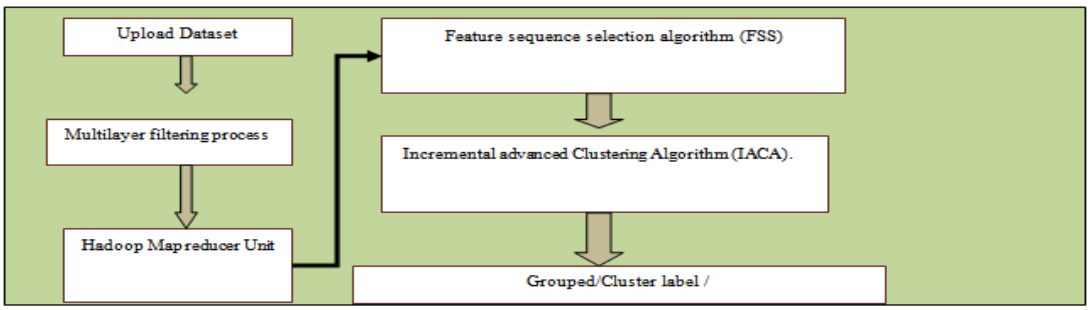


Fig 3.1 Proposed system architecture

Contribution Of The Proposed Work

The followings are the contributions of the proposed system.

- The previous framework is one of the iterative frameworks and it requires repeated data re-clustering with an incrementally growing constraint set. In this set can be computationally demanding for large data sets with high dimensional. This problem is addressed then the system introduces an incremental advanced Clustering Algorithm method that updates the existing clustering solution based on the neighborhood assignment for the new point.
- An alternative way to lower the computational cost is to reduce the number of iterations by applying a Feature sequence selection this process select effective feature of the dataset.

IV. RESEARCH METHODOLOGY

Basically, the clustering is classifies or partition the given data set, so it means technically "the act of partitioning an unlabeled dataset into groups of similar objects". The clustering main goal is to group sets of objects into classes such that similar objects are placed in the same cluster while dissimilar objects are in separate clusters. Then the proposed system performs the semi supervised clustering.

1. Multilayer filtering,
2. Hadoop map reducer,
3. Feature sequence selection,
4. IACA Algorithm.

Multilayer Filtering

The data that is collected is in file format. The data in the real world is highly susceptible to noise containing errors or outliers, missing, and inconsistency. Therefore, pre-processing of data is very important. Multilayer Filtering algorithm has been applied to effective pre-processing. This process is implemented in some steps as follows:

Step 1: Import the libraries
Step 2: Import the data-set
Step 3: Check out the missing values
Step 4: "Discretization" which is unsupervised attribute filters changes numeric data into nominal
Step 5: process continue until total dataset length
Step 6: Feature Scaling

Hadoop Map reducer

The term "**Map Reduce"** is a programming model. It has used for processing the large data sets with a parallel and distributed algorithm on a cluster. The basic unit of information, used in Map Reduce is a (Key, value) pair. All types of structured and unstructured data need to be translated to this basic unit, before feeding the data to Map Reduce model. Map Reduce model consist of two separate routines, namely Map-function and Reduce-function.

Feature Sequence Selection

Feature selection is a process commonly used in the second stage of the proposed work, wherein a subset of the features available from the data is selected for application of a learning algorithm. Although the algorithmic term selects the best subset and that it contains the least number of dimensions that most contribute to accuracy. This discards the remaining, unimportant dimensions using feature sequence selection algorithm. This is an important stage after pre-processing and is one of two ways of avoiding the curse of dimensionality in the health care domain. There are two approaches in feature selection namely forward selection and backward selection. Forward Selection starts with no variables and adds them one by one, at each step adding the one that decreases the error the most, until any further addition does not significantly decrease the error. Backward Selection starts with all the variables and removes them one by one, at each step removing the one that decreases the error the most, until any further removal increases the error significantly.

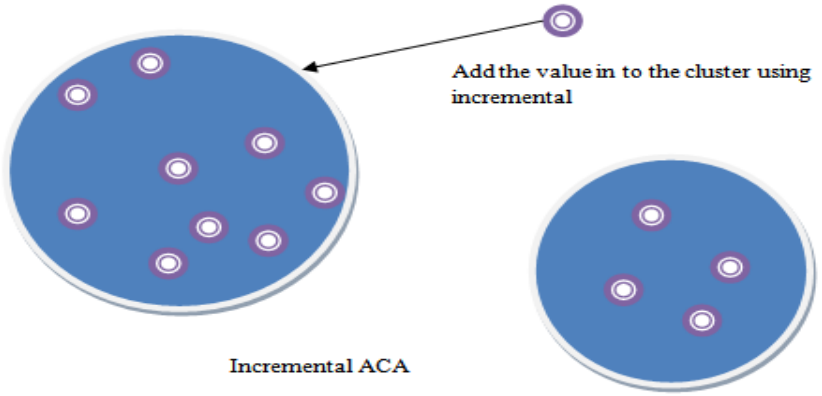
Algorithm Steps

Input: D (F0, F1, ..., Fn−1) // a data set with list of N features
S0 // a subset from which to begin the search
δ 1// a stopping criterion
Output: Sbest // an optimal subset
Step 1: begin initialize: Sbest = S0;
Step 2: γbest = eval (S0, D, A); // evaluate S0
Step 3: S = generate (D); // generate a subset for evaluation

Step 4: **γ1** = eval(S, D, A); // evaluate the current subset S by A
Step5: if (γ1 is better than γbest)
Γbest = γ1;
Sbest = S; //assign
Step 6: End until (δ is reached);
Return Sbest;

IAC Algorithm

The system effectively utilizes the incremental advanced Clustering Algorithm. But this algorithm has not only great robustness, because it gives the positive feedback characteristic and also with parallel and distributed computing feature.



V. RESULTS AND ANALYSIS

The experiments are basically designed so that the different parts of the work could be evaluated easily and effectively. To this aim, first the features which were selected by the feature selection method named as feature sequence selection and their importance are discussed. Second, all the four possible combinations of the feature selection and creation methods are theoretically analyzed over the heart disease dataset completely. Finally algorithms are implemented this proposed work was implemented using Java. The performance of this proposed work Scheme was compared with the existing algorithms based on the following parameters.

Data Sets

In the experimental is uses benchmark data set from the UCI repository. It which not has been used in previous studies on constraint based clustering. Then the term heart disease includes the datasets. So, it can be apply the new techniques and processed the related dataset.

Dataset Description:

Benchmark UCI data sets:
Dataset Name: heart disease
URL: http:\\www.UCIrepostory.org\statlog_heart

Table 5.0 Comparison Table

Name	Type	Description
Age	Continuous	Age in years
Sec	Discrete	1=male 0=female
Cp	Discrete	Chest pain type: 1=typical angina 2=atypical angina 3=non-anginal pa 4=asymptomatic
Trestbbs	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar>120 mg/dl: 1=true 0=false

Restecg	Discrete	Resting electrocardiographic results: 0=normal 1=having ST-T wave abnormality 2=showing probable or define left ventricular hypertrophy Estes Criteria
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: 1=yes 0=no
Slope	Discrete	The slope of the peak exercise segment: 1=up sloping 2=flat 3=down sloping
Diagnosis	Discrete	Diagnosis classes: 0=healthy

The data set has 13 attributes. The experiment takes Heart disease dataset from UCI repository. The dataset contains 13 attributes considered are: age, sex, FBS (fasting blood sugar > 120 mg/dl), chol (serum cholesterol in mg/dl), restecg (resting electrocardiographic results), trestbps (resting blood pressure), thalach (maximum heart rate achieved), exang (exercise induced angina), slope (the slope of the peak exercise ST segment), oldpeak (ST depression induced by exercise relative to rest). There are a total of 250 patient records in the database.

Experimental Results

This section describes the implementation process. Implementation is the realization of an application, or execution of plan, idea, model, design of a research. This section explains the software, datasets and modules which are used to develop the research. Then experimental term is performed on an Intel Dual Core with a RAM capacity 2GB. The algorithms are implemented in Java and are run under Windows platform.

- **Specificity** –measures the proportion of negatives that are correctly identified.
- **Sensitivity**- measures the proportion of positives that are correctly identified
- **Accuracy** – Determines the correctness
- **Precision** –Repeated process same result
- **Time taken** – Determines the processing time involved.

TP, TN, FN and FP these terms are described by Sensitivity, specificity and accuracy.

Performance Evaluation

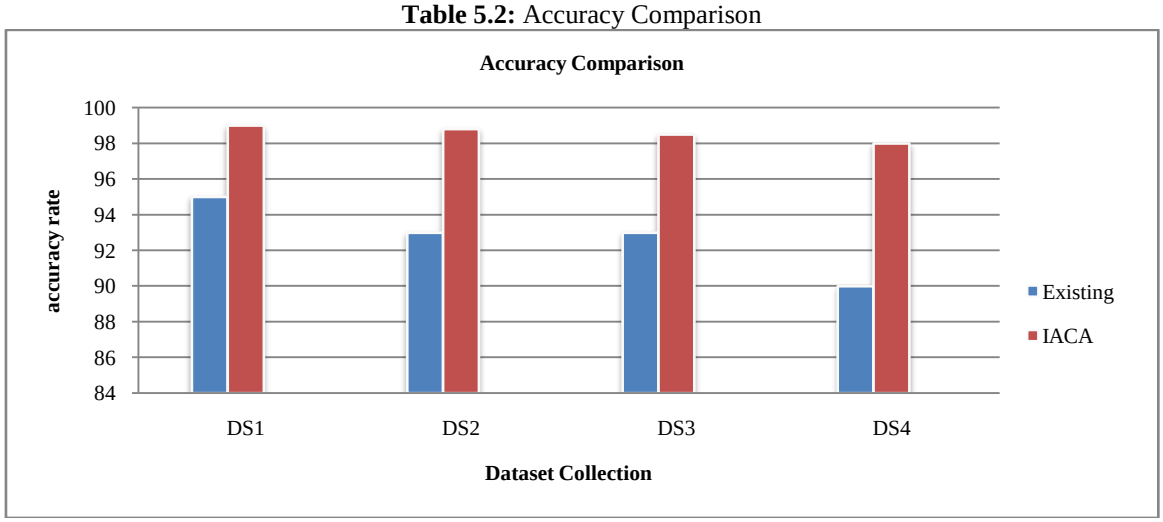
This experiment has been done through the heart disease Dataset. The dataset is preprocessed by Multilayer filtering and features are selected using FSS and finally the clustering process is made BY IACA.). The proposed system detects objects and its classes with significant improvement in terms of high classification accuracy this can be analyzed with different set of data's and results are shown

Table 5.1 Performance Evaluation

Metrics	**Dataset**	**Existing**	**Proposed IACA**
	DS1(100)	95	99
Detection Accuracy (%)	**DS2(150)**	93	98.8
	DS3(200)	93	98.5
	DS4(250)	90	98

Performance comparison of proposed IACA with existing approaches based On Heart Disease Detection Result accuracy

Table 5.2: Accuracy Comparison



The false positive rate of the proposed system is quite high, because some normal classes in the additional data merged could be clustered as a disease, but only the weighted features are used in grouping. The reduction in false positive rate of the proposed system is mainly due to the IACA process.

Table: 5.3 Performance comparison table

Process	**Existing**	Proposed system
Pre-Processing time	2.8	1.5
Feature selection Time(s)	8	3.4
Clustering Time (s)	5	2.6

Precision: A class is the number of true positives (i.e. the number of instances correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class) that is called "**Precision**". The equation is:

Precision = TP / (TP + FP)

Recall: The Context of "**recall"** is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. to take true positives and false negatives and sum these values, which are items which were not labeled as belonging to the positive class. But it should have been.) The Recall can be calculated as:

Recall = TP / (TP + FN)

Accuracy: The percentages of the predicted values are match with the expected value for the given data. The advantages like, high Accuracy, High Precision and High Recall value is give which system that is system is the best system to be considered it. The performance of the proposed system is tested with the 5000 instances, from each instance the precision and recall values are gathered and that is plotted in the fig 4.5. With help of the confusion matrix from table 4.3values measurement of the precision and recall values are calculated and plotted as a graph below:

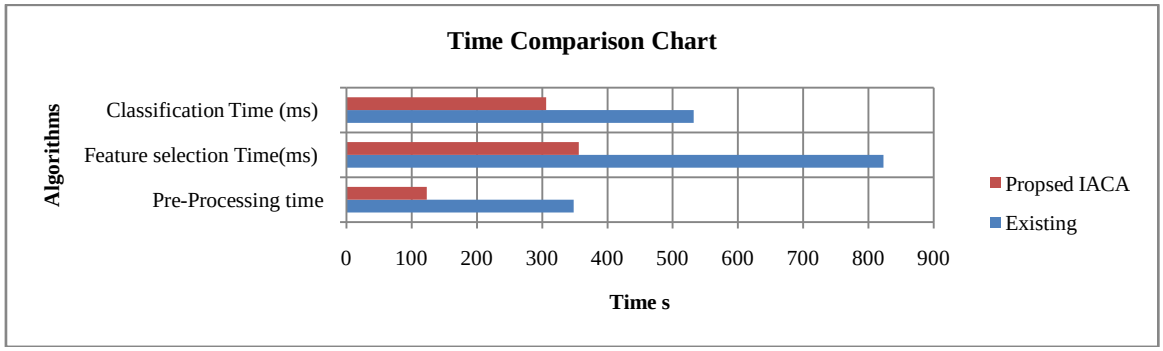


Fig: 5.4 Time comparison between existing and proposed

It observed that the performance is very promising compared to the existing methods that have been explored in the previous chapter. The next chapter deals with the presentation of the conclusion and enhancements.

VI. CONCLUSION

The study proposed a new clustering and prediction scheme for Heart disease data. The system studied the main two problems in the literature, which are detection accuracy and delay. The study overcomes the above two problem by applying the effective enhanced IACA algorithm. The proposed system handles the large category dataset more rapidly, accurately and effectively. It keeps the good scalability at the same time. The system effectively labels the object cluster which is referred such as normal and disease. The experimental results are evaluated using the Java. The experimental result shows that integrated extended proposed algorithm shows better quality assessment compared to traditional clustering techniques. From the experimental results, the execution time calculated for clustering object label is almost reduced than the existing system.

REFERENCES

[1]. S. Sharmila. "Analysis of heart disease prediction using data mining technique", JCSE, 2017.
[2]. Aqueel Ahmed, Shaikh Abdul Hannan, **"**Data Mining Techniques to Find Out Heart Diseases: An Overview", International Journal of Innovative Technology and Exploring Engineering (IJITEE), September 2012.
[3]. Ramin Assari, Parham Azimi, Mohammad Reza Taghva. "Heart Disease Diagnosis Using Data Mining Techniques", Int J Econ Manag Sci, ISSN: 2162-6359, Volume 6 • Issue 3 • 1000415,2017.
[4]. Mudasir M Kirmani. "Cardiovascular Disease Prediction Using Data Mining Techniques: A Review". Orient.J. Comp. Sci. and Technol;10(2) ,June, 2017.
[5]. Aditya Methaila, Prince Kansal,Himanshu Arya, Pankaj Kumar. "Early Heart Disease Prediction Using Data Mining Techniques". CCSEIT, DMDB, ICBB, MoWiN, AIAP – 2014.

Susmitha K"A Novel Machine Learning Approach to Predictions in Heart Disease Using Iaca""IOSR Journal of Engineering (IOSRJEN), vol. 08, no. 12, 2018, pp. 67-73