RESEARCH ARTICLE OPEN ACCESS

Classification Algorithms: A Recent Empirical Comparative Survey

Sapna Sharma*, Dr. Pankaj Sharma **
**(Department of Computer Science and Technology, Sachdeva Institute of Technology Mathura)*
*** (Department of Computer Science and Technology, Sachdeva Institute of Technology Mathura)*
Corresponding Author: Sapna Sharma

ABSTRACT
Healthcare industry collects large amounts of clinical data of patients. Although this data is useful but not provides hidden information for effective decision making. Advanced data mining techniques can be used to discover hidden pattern for effective decision making. Medicinal data mining methods are used to analyze the medical data. Medical data mining content mining and structure methods are used to analyze the medical data contents. Diagnosis of heart disease is a significant and tedious task in medicine. The term Heart disease encompasses the various diseases that affect the heart. There are several classification techniques like Naïve Bayes classification, Support Vector Machine, Decision Tree classifier, Artificial neural networks (ANNs) classifier, and Rule based classifiers. Performance of these techniques is compared through sensitivity, specificity, accuracy, error rate, True Positive Rate and False Positive Rate. In this paper we proposed a comparative study of recent classification techniques on the basis of some parameters
Keywords - Classification, Prediction, Accuracy, Diagnosis, Heart

---

Date Of Submission: 05-08-2019 Date Of Acceptance: 20-08-2019

I. INTRODUCTION

The diagnosis of diseases is a significant and tedious task in Medical science. Most hospitals today employ sort of hospital information systems to manage their healthcare or patient data. How data is turned into useful information that can enable healthcare practitioners to make intelligent clinical decisions. Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. The exposure of heart disease from various factors or symptom is an issue which is not complimentary from false presumptions often accompanied by unpredictable effects.

II. CLASSIFICATION

Classification is an one of the most important data mining techniques. Classification involve in the procedure of finding a model to distinguish data classes. Classification classifies objects into one of the predefined classes or groups. Classification techniques used various mathematical models such as linear programming, regression and statistics. The aim classification model is to divides data objects into unknown target classes. For example, giving loan to a person can be classified as "Yes" or "No" on the basis of their credit rating using data classification techniques. There are two types of method used in the classification.

1. Binary classification
2. Multilevel classification

In binary classification there are only two possible classes such as, "Yes" or "No". In multiclass there are more than two classes for example patient can be "high", "medium" and "low" risk patient.

Classification is a two step process

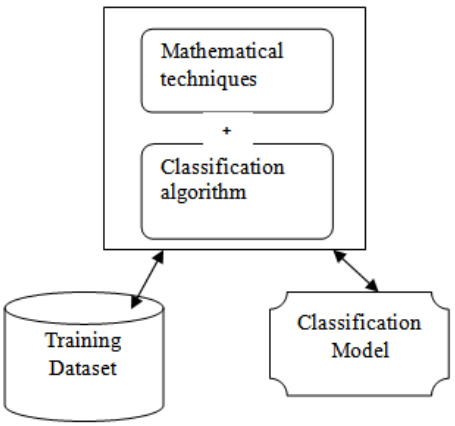


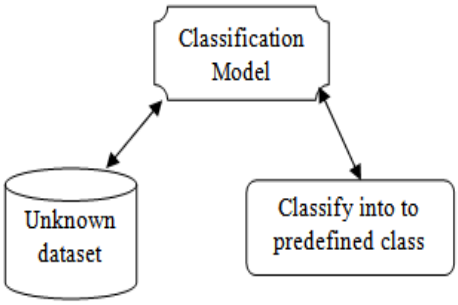
Figure 1 constructing classification model

1. Learning Phase:

In this phase construct of classification model by using different algorithms. The model then used to build a classifier by using training set. Accuracy of the model has tested against various training dataset for good accuracy

2. Prediction Unknown Tuple

Model constructed in the first phase used to predict unknown class. Estimate the accuracy of unknown tuple by checking correct class label.



III. TECHNIQUES OF CLASSIFICATION

There are several techniques have been developed in the past year, each technique used different concepts and formula to classify the given dataset. Some of the most important techniques which are used commonly are given below

1.Decision Tree classifier

In early 1980, J. Ross Quinlan developed a decision tree algorithm In 1984 L. Breiman developed Classification and Regression Trees which described the generation of binary decision trees. Decision tree induction follows a top-down approach. A decision tree is a flowchart-like tree structure, in which each internal node presents a test for attribute, branch shows outcome for the test, and each terminal node belongs to class label. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy.

2 Naïve Bayesian Classification

Bayesian classifier is statistical classifiers and able predict class membership probabilities. Bayesian classification is based on Bayes' theorem. Bayesian classifiers have good accuracy and speed when applied for large databases.

Bayesian classifier works as follows:

Let D be a training set. D is collections of tuple represented by dimensional attribute vector, $\mathbf{X} = (x_1, x_2, : : : , x_n)$ with n attributes $A_1, A_2, \ldots\ldots\ldots, A_n$.

2. $C_1, C_2, \ldots\ldots, C_m$ represent classes for given a tuple, classifiers predict **X** belongs to the class having the highest posterior probability

Bayesian classifier predicts that tuple **X** belongs to the class C_i if and only if

$P(\mathbf{C}_i|\mathbf{X}) > P(\mathbf{C}_j|\mathbf{X})$ for $1 \leq j \leq m; j \neq i$

3.Rule-Based Classification

Rules are a good way of representing knowledge. A rule-based classifier is based on some set of rules. A rule R1 is denoted in the form of

IF condition THEN conclusion

For example R1: IF age = youth AND govt. employee = yes THEN loan = yes.

In the rule the condition consists of one or more attribute tests that are logically AND.

If the condition in a rule holds true for a given tuple, then the rule satisfied and the rule covers the tuple.

4. Back propagation as classifier

Back propagation is based on neural network learning. A neural network is a set of connected input/output units in which each connection has a weight. The neural network learns by adjusting the weights to predict the correct class label. Neural networks have high training times and require a number of parameters to determine empirically. Advantages of neural networks include their high tolerance of noisy. Neural network are useful even if we have little knowledge of the relationships between attributes and classes.

5. Support Vector Machine

Support Vector Machines was presented in 1992 by Vladimir Vapnik. Support Vector Machines classify both linear and nonlinear data. Support Vector Machine uses a nonlinear mapping to transform the original training data into a higher dimension; it searches for the linear optimal separating hyper plane. The support vectors also provide a compact description of the learned model. SVMs are useful for both prediction as well as classification. SVM have been applied to a number of areas, including handwritten digit recognition, object recognition, and speaker identification, as well as benchmark time-series prediction tests.

6. Genetic Algorithm

Genetic algorithms are used as classifiers. Genetic learning starts with initial population consisting randomly generated rules. Rules are represented using string of bits. For example for a given training set two attributes, A_1 and A_2, and that there are two classes,C_1 andC_2. The rule"IF A_1 AND NOT A_2 THEN C_2" are encoded using bit string "100". Similarly, the rule "IF NOT A_1 AND NOT A_2 THEN C_1" can be encoded as "001". Typically, the fitness of a rule is assessed by its classification accuracy on a set of training samples. In crossover, substrings from pairs of rules are swapped to form new pairs of rules.

IV. LITERATURE REVIEW

In 2010 Sunita Soni et al. proposed "Associative Classifiers for Predictive Analysis in Health Care Data Mining". They used a combined approach that integrates association rule mining and classification rule mining. They also introduce that combining the advanced association rule mining with classifiers gives a new type of associative classifiers with small refinement in the definition of support and confidence that satisfies the validation of downward closure property [1].

In 2010 N. Suneetha et al proposed "Modified Gini Index Classification". They proposed a tree based approaches. They analyzed multiple response using classification algorithms comparing with a modified decision tree method. They normalized the Gini indexes by taking into account information about the splitting status of all attributes. Instead of using the Gini index for attribute selection as usual, they used ratios of Gini indexes and their splitting values in order to reduce the biases [2].

In 2011 G. Subbalakshmi et al. proposed "Decision Support in Heart Disease Prediction System using Naive Bayes". They proposed a Decision Support in Heart Disease Prediction System (DSHDPS) using data mining modeling technique, namely. They implement the system by using web based questionnaire application. This system helps to train nurses and medical students to diagnose patients with heart disease. Decision Support in Heart Disease Prediction System is developed using Naive Bayesian Classification technique [3].

In 2011 Mai Shouman et al. proposed "Using Decision Tree for Diagnosing Heart Disease Patients". They investigates and applying a range of techniques to different types of Decision Trees seeking better performance in heart disease diagnosis. A widely used benchmark data set is used in this research. To evaluate the performance of the alternative Decision Trees the sensitivity, specificity, and accuracy are calculated. The research proposes a model that outperforms J4.8 Decision Tree and Bagging algorithm in the diagnosis of heart disease patients. [4].

In 2012 M. Akhil Jabbar et al. proposed "Heart Disease Prediction System using Associative Classification and Genetic Algorithm". They proposed an efficient associative classification algorithm using genetic approach for heart disease prediction. They proposed a system for heart disease prediction using data mining techniques[5].

In 2012 Chaitrali S. et al. proposed "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques". They analyzed prediction systems for Heart disease using more number of input attributes. The system uses medical terms such as sex, blood pressure, cholesterol like 13 attributes to predict the likelihood of patient getting a Heart disease. Until now, 13 attributes are used for prediction. They added two more attributes i.e. obesity and smoking. The performance of these techniques is compared, based on accuracy [6].

In 2012 Sunita Soni et al "Fuzzy Weighted Associative Classifier: A Predictive Technique For Health Care Data Mining". They extend the problem of classification using Fuzzy Association Rule Mining and propose the concept of Fuzzy Weighted Associative Classifier (FWAC). Classification based on Association rules is considered to be effective and advantageous in many cases. They proposed a new Fuzzy Weighted Associative Classifier (FWAC) that generates classification rules using Fuzzy Weighted Support and Confidence framework. [7].

In 2013 M. Akhil Jabbar et al. proposed "Heart Disease Classification Using Nearest Neighbor Classifier with Feature Subset Selection". They investigate and apply K nearest neighbor with feature subset selection in the diagnosis of heart disease. The experimental results show that applying feature subset selection to KNN will enhance the accuracy in the diagnosis of heart disease for Andhra Pradesh population. India with a population of more than 1 billion accounted for 60%of the world heart diseases. [8].

In 2013 M. Akhil Jabbar et al proposed "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection". They introduced a classification approach which uses ANN and feature subset selection for the classification of heart disease. PCA is used for preprocessing and to reduce no. Of attributes which indirectly reduces the no. of diagnosis tests which are needed to be taken by a patient. They proposed a new feature selection method for heart disease classification using ANN and various feature selection methods for Andhra Pradesh Population.[9].

In 2014 Mariammal D et al. proposed "Major Disease Diagnosis and Treatment Suggestion System Using Data Mining Techniques". They proposed a model to systematically close those gaps to discover if applying single and multiple data mining techniques to all disease treatment data can provide as reliable performance as that achieved in diagnosing disease. Using multiple data mining techniques the accuracy also improved. Instead of going for a number of tests, predicting the major disease with less number of attributes is a challenging task in Data Mining. [10].

In 2015 Ebenezer Obaloluwa et al. proposed "Heart Diseases Diagnosis Using Neural Networks Arbitration". They proposed causes of

heart diseases, the complications and the remedies for the diseases have been considered. An intelligent system which can diagnose heart diseases has been implemented. The dataset of heart disease has been used to carry out this experiment. The dataset comprises attributes of patients diagnosed for heart diseases. The diagnosis was used to confirm whether heart disease is present or absent in the patient [11].

In 2016 Isra'a Ahmed et al. proposed "A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods". They motivate is to develop an effective intelligent medical decision support system based on data mining techniques. They used five data mining classifying algorithms, with large datasets, have been utilized to assess and analyze the risk factors statistically related` to heart diseases in order to compare the performance of the implemented classifiers. [12].

In 2017 Sanjay Kumar Sen proposed "Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms". They carried out an experiment to find the predictive performance of different classifiers. They select four popular classifiers considering their qualitative performance for the experiment. They also choose one dataset from heart available at UCI machine learning repository. Naïve base classifier is the best in performance [13].

In 2018 Poornima V et al proposed "A novel approach for diagnosing heart disease with hybrid classifier". They proposed an Orthogonal Local Preserving Projection (OLPP) method to reduce the function dimension of the input high-dimensional data. The final output of the optimization technique is combined with the performance metrics as accuracy, sensitivity, and specificity. From the result, it is observed that hybrid optimization techniques increase the accuracy of the heart disease prediction system [14].

V. PROBLEM STATEMENT

There are various classification techniques that can be used for the identification and prevention of heart disease. The performance of classification techniques depends on the type of dataset that we have taken for doing experiment. Classification techniques provide benefit to all the people like doctors, patients and organizations who are engaged in healthcare industry. Decision tree, Bays Naive classification, Support Vector Machine, Rule based classification, Neural Network as a classifier etc. The main problem related to classification techniques are

1) **Accuracy: -** This includes accuracy of the classifier in term of predicting the class label, guessing value of predicted attributes.
2) **Speed:-**This include the required time to construct the model (training time) and time to use the model (classification/prediction time)
3) **Robustness:-**This is the characteristic of the classifier or predictor to make correct prediction and give correct result on noisy data.
4) **Scalability:-**Efficiency in term of database size.

VI. COMPARATIVE ANALYSIS

S. N	Tech	**Benefits**	**Limitations**
1	K-NN	It is easy to implement.	Sensitive to noise.
2	Decision Tree	It can easily process high dimension data	It is restricted to one output attribute. .
3	SVM	Better Accuracy	Computationally expensive.
4	Neural Network	Easily identify complex relationships	Local minima. and Over-fitting.
5	Bayesian	Computations process easier.	It does not give accurate results where dependency exist among variables

VII. CONCLUSION

There are several techniques have been developed in the last decades but efficiency and improvement and improvement are always required, so there is always required new method and techniques. Developing new method and techniques is always a research area. In this paper we proposed detail study of existing classification techniques.

REFERENCES

[1]. Sunita Soni O. P. Vyas Using Associative Classifiers for Predictive Analysis in Health Care Data Mining International Journal of Computer Applications (0975 – 8887) Volume 4 No.5, July 2010.
[2]. N. Suneetha CH.V. M. K. Hari and V. Sunil Kumar Modified Gini Index Classification: A Case Study Of Heart Disease Dataset (IJCSE) International Journal on Computer Science

and Engineering Vol. 02, No. 06, 2010, 1959-1965.

[3]. Mrs. G. Subbalakshmi and Mr. K. Ramesh Decision Support in Heart Disease Prediction System using Naive Bayes Journal of Computer Science and Engineering (IJCSE) ISSN : 0976-5166 Vol. 2 No. 2 Apr-May 2011.

[4]. Mai Shouman, Tim Turner, Rob Stocker Using Decision Tree for Diagnosing Heart Disease Patients Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia Copyright © 2011, Australian Computer Society

[5]. M. Akhil jabbar, Dr B.L Deekshatulu, Dr. Priti Chandra "Heart Disease Classification Using Nearest Neighbor Classifier With Feature Subset Selection Computer Science and Telecommunications 2013|No.3(39) ISSN 1512-1232.

[6]. Chaitrali S. Dangare Sulabha S. Apte, " Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012.

[7]. Sunita Soni and O.P.Vyas "Fuzzy Weighted Associative Classifier: A Predictive Technique For Health Care Data mining" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.1, February 2012.

[8]. M. Akhil jabbar, Dr B.L Deekshatulu, Dr Priti Chandra3 " Heart Disease Classification Using Nearest Neighbor Classifier With Feature Subset Selection" GESJ: Computer ScienceandTelecommunications 2013|No.3(39).

[9]. M. Akhil Jabbar, B.L Deekshatulu & Priti Chandra "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection" Global Journal of Computer Science and TechnologyNeural & Artificial Intelligence Volume 13 Issue 3 Version 1.0 Year 2013 Type: Double Blind Peer Reviewed International Research Journal Publisher: Global Journals Inc. (USA) Online ISSN: 0975-4172 & Print ISSN: 0975-4350.

[10]. Mariammal. D, Jayanthi. S, Dr. P. S. K. Patra Major Disease Diagnosis and Treatment Suggestion System Using Data Mining Techniques International Journal of Advanced Research in Computer Science & Technology IJARCST All Rights Reserved 338 Vol. 2 Issue Special 1 Jan-March 2014 ISSN: 2347 - 8446 (Online) ISSN: 2347 – 9817.

[11]. Ebenezer Obaloluwa Olaniyi and Oyebade Kayode Oyedotun "Heart Diseases Diagnosis Using Neural Networks Arbitration" I.J. Intelligent Systems and Applications, 2015, 12, 75-82 Published Online November 2015 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijisa.2015.12.08.

[12]. Isra'a Ahmed Zriqat, Ahmad Mousa Altamimi, Mohammad Azzeh "A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods" International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 12, December 2016 ttps://sites.google.com/site/ijcsis/ ISSN 1947-5500.

[13]. Sanjay Kumar Sen "Predicting and Diagnosing of Heart Disease Using Machine Learning Algorith" International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 6 Issue 6 June 2017, Page No. 21623-21631 Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i6.14.

[14]. Poornima V, Gladis D A novel approach for diagnosing heart disease with hybrid Biomedical Research 2018; 29 (11): 2274-2280 ISSN 0970-938X www.biomedres.info

Sapna Sharma" Classification Algorithms: A Recent Empirical Comparative Survey" International Journal of Engineering Research and Applications (IJERA), Vol. 09, No.08, 2019, pp. 69-73