International Journal of Electrical Electronics & Computer Science Engineering
Special Issue - TeLMISR 2015, ISSN : 2348-2273
Available Online at www.ijeecse.com

An Analytical Study on Classification Algorithms for Medical Datasets

Alpna Sharma, Seema Sharma
VSIT, Vivekananda Institute of Professional Studies, Delhi, India
alpna81@gmail.com, seemasharma08@ymail.com

Abstract: ***Medical Information Processing is having its significance to perform the effective disease recognition over the dataset. There are number of available algorithms present to perform the recognition. But there is the requirement to identify the most reliable and accurate method. In this work, an analytical study on some of effective recognition methods is defined. Here four effective learning approaches are analyzed on four medical datasets. These approaches are decision tree, random tree, decision table and Bayesian network approach. The results shows that the Bayesian network is most effective and reliable approach among all these methods.***

Keywords: Decision tree, Decision Table, Bayesian, Medical Data Processing

I. INTRODUCTION

One of the major aspect of data mining is classification algorithm. These algorithms are been used categorized the data among defined number of classes based on feature extraction. The classification is one of the core application required in each domain either it is medical disease prediction, intrusion detection, spammers identification etc. The classification process is defined under some machine learning assisted approaches with rule specifications. These rules help to map the data to various classes. There are number of associated classification approaches shown in figure 1. These approaches are divided in supervised and unsupervised classification approaches. The classification problem is defined with the specification of object need with the attribute as well as value level analysis. These classification methods uses a set of features that are characterize with other objects and generate the relevant information to take the conclusion about the associated class. To perform the classification, the dataset is divided in terms of training and testing methods. In this work, a study based work is defined on various classification approaches. This classification is here applied on various medical datasets. In third section, the description about the classification algorithms and the datasets is given.

The classification process is actually defined as a prediction or recognition system in which the hidden features from the dataset are extract to take the relevant decisions. The main work is here defined to take the effective decisions to obtain the information from the pattern analysis derived from the dataset. This knowledge extraction is here been doing under the specification of contributing factors. These all factors are defined with the specification of data range and the transient data specification with the help of effective algorithms as well as the automated tools. These algorithms or tools processed the primary data in an intelligent way perform various levels of transformation so that the decision adaptive information is obtained from the dataset. The explosive information growth is then defined with specification of the automation information derivation tool as well as to perform the information discovery.

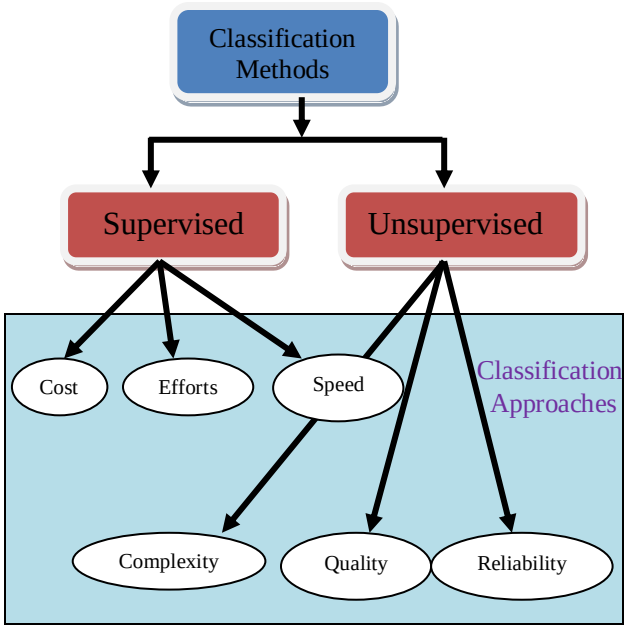

Figure 1 : Classification Approaches

A) Classification In Medical Field: Medical information process is one of the critical and responsive area for classification and recognition algorithms. This kind of recognition system comes under the intelligent data mining and expert system. The quality of the obtained result must be high for these application areas. This prediction systems are defined respective to particular disease and provide the identification under the specification of set of associated rules. The identification of trends or the patterns from the dataset for a specific medical disease is also a challenge. This kind of

recognition systems are based on the patient symptoms as well as the environmental and situational vectors. The reliability of the dataset is also a key issue in such kind of recognition system. The data mining performed in this area is required to test under various vectors that are tested on different patient samples. Even after all kind of results are presented under the guidance of some physician who is expert in that area. In this paper, some of the medical diseases are considered for the analytical prediction. These disease includes heart disease, diabetics, hepatitis etc.

In this paper, an analytical study is been presented on medical datasets to perform the disease classification based on feature analysis. The work is here tested on four different classification methods respective to four datasets. The analysis of work is here done under multiple vectors. In this section, the classification approach is been discussed along with the categorization. The section also explored the process of medical prediction system. In section II, the work defined by earlier researchers is discussed. In section III, the study based work is defined on four classification algorithms. In section IV, the results obtained from the work are presented. In section V, the conclusion derived from the work is presented.

II. EXISTING WORK

Lot of work is already defined by various researchers on medical classification algorithm. Some of the work defined by different researchers on different medical domains and approaches is presented in this section. Jesmin Nahar[1] has defined a work on cancer risk assessment under different vectors based on rule discovery. Author has analyzed the work under different risk factors and presented a study on skin cancer, lung, breast cancer etc. Author studied various mining approaches respective to different algorithmic approaches. These approaches includes risk factors includes apriori methods, predictive methods and tertius algorithm. Author has defined the confidence value analysis for the effective prediction under confidence value so that the rule effective decisions will be taken. Author studied the significant risk factors to analyze the risk over the dataset. Carlos Ordonez[2] has defined a work on association rule discovery on medical processes. Author has defined a work on association rule mining under constraint analysis. The work is applied on heart disease prediction. The prediction is based on transactional rule mapping with derivation of association rules and identification of the constraints over it. The constraint specification is here been defined with the specification of rules associated with dataset. Adepele Olukunle[3] has defined medical image data processing under association rule specification. Author defined a work on fast association information processing with specification of mining rules. Author defined a work on feasibility study under association rule algorithm and to extract the information so that the hidden information will be derived . Author defined a falour of implementation so that the information suitability will be obtained. Carlos Ordonez [4] has defined a work based on disease prediction applied on heart disease dataset. Author defined a search constraint analysis under rule specification and search with training rule generation and validating under independent test set. Author defined the work under the significance of discovered rules and evaluation with support, confidence and lift values. Association rules are applied for dataset deviation and heart perfusion constraint specification so that the significant results will be obtained. The dataset validations are here been defined so that the reduced number of association rules are defined with predictive accuracy. Author defined the high confidence, high lift and valid information processing. Author defined medical information discovery over the dataset.

Masaya Yoshikawa[5] has defined a work on ant colony optimization with job shop scheduling problem. Author defined a work on the optimization approach so that the effective scheduling will be performed and the prediction accuracy will be obtained. Author defined the work on the scheduling problems so that the comparative decisions will be taken. Shen Xiaoyan[6] has defined a work on differentiation oriented work for disease prediction under association rule mining. Author defined the frequent information processing with item set processing with dimension processing and reduction. Author used the growth algorithms for the prediction and reduction of dataset. Laila Elfangary [7] has defined a work on hidden pattern discovery applied on large medical dataset. The work is defined to derive the normal results from conventional techniques. Author defiend a work on symptom and diagnosis processing with the specification of feature vectors such as rules, classification, clarity, automation etc. Wenzhi Zhu[8] has defined a work on recognition system using ACO approach. Author defined the fitness function with the control specification rule so that the steady state error will be reduced. NING Hong-yun[9] has defined a work on rule chain specification based mining approach for potential association mining under directed graph. Author defined the rule chain specification with dynamic heuristic feature and intensity information analysis. The feedback and pheromone analysis is been defined with higher selection probability. Haiwei Pan[10] has defined a work on knowledge constraint under domain information processing. Author defined a stage level work to extract the information

based on association rule and the formed the medical information processing under rule level extraction.

Gaurav N. Pradhan[11] has defined work based on mult dimensional timer series based medical data. Author has analyzed the muscular activities of human participants and obtained the EMG decisions for pattern discovery so that the multiple information decisions will be taken. The information patterns are extracted over the time line and generic decisions are taken to improve the recognition process. Kittisak Kerdprasop[12] has defined a work to manage the medical datasets and perform the knowledge information mining in an integrated form. Author processed the knowledge discovery process under semi automatic triggering so that the predictive prototyping decisions will be taken. Zhongmei Zhou[13] has defined a work on association specification so that the probabilistic decision will be taken regarding the medical information processing. Author applied on medical dataset and derived the effective results.

III. CLASIFICATION ALGORITHMS

In this section, the predictive data mining is performed on medical datasets using four different classification algorithms. These algorithms includes Bayesian Network, Decision Tree, Decision Table and Random Tree Random Tree Algorithm. In this section, a descriptive approach is defined for these all algorithms. All these algorithms are having some common model for the processing. According to this model, the dataset is divided in training and testing sets and the training sets are having the results vectors in classification vectors and the testing set is considered as the input set on which the recognition will be applied. The description of these classification algorithms is given here under

A) Bayesian Network: Bayesian Network is defined a classification model based on the concept of conditional probability. Here the probabilistic relation analysis is been defined among various variables. The Bayesian network is here defined under belief network specification under directed acyclic model in which the conditional dependencies are defined with each node. The classifier learns are been defined under training data specification for each attribute and derived to the related class. The classification bayes rules are defined to estimate the probability under the instance specification. The predictive class derivation with posterior probability is defined to take the decision under attribute level estimation. The correspondence with the feature extraction is been done to derive the specification on conditional independencies. The rule specification is here been done under the formation. The algorithmic specification of work is given here under

- The dataset is divided in the form of training set with specification of class label and the attribute class is defined with N attribute vectors.
- The m classes are defined for recognition
- The classification process is defined along with maximum posteriori specification
- Bayes theorem is defined to perform the recognition
- The probabilistic class is defined for derive the result.

B) Decision Tree: Decision Tree is defined as the learning model based on the statistical analysis and machine learning rules. The predictive model is defined to map the observations and to generate the tree learning specification. The predictive model is defined under specification of input set and the target values. The descriptive tree model is defined under structure level analysis with conjunctive feature specification respective to the instance and the root node. The feature analysis with generalization is defined to obtain the information gain. The decision level partitions are defined to obtain the data point generation and transparent structure generation with internal and external node connectivity analysis. The derivation is ere been defined with specification of relevant component and decision criteria. The algorithmic representation is given here under.

- Generation of Node for Decision Tree
- The sample set analysis is defined with classification
- The sample is analyzed respective to the class and identify the respective disease class
- The test attribute is done under information gain analysis for test attribute analysis.
- The class derivation is performed for generation of attribute.

C) Random Tree: The random tree is the improved form of decision tree that defines the random rule set based on predictive analysis. The specification of rule is here defined under dissimilarity analysis and instance analysis under feature level evaluation. The specification of the information gain under the data instance and ambiguity analysis. Author defined an instance falling category based on target value analysis. Author defined a work on the target based estimation applied on each branch and the branch value estimated based on targeted value. The specification of the decisions is here been done under the attribute level analysis and derive the item value analysis.

These probabilistic rules will be applied at each level to take the class level decision. When all node level decisions are merged over all decision about the recognition or classification process is defined.

D) Decision Table: It is one of the effective and simplest learning method comes under supervised learning. It uses the decision rules in the form of a table and perform distance level analysis to identify the appropriate class. The training and testing both vectors and based on the statistical measure and the relatively variety of decisions are taken. The statistical decision learning algorithms are here defined under variation analysis so that the effective prediction of disease over the dataset will be done. The approach is able to convert an unknown point to known point vector so that the distance metric will be obtained under training vector. Based on these point level parameters the instance learning is applied and the decision about the associated class is done. The work is about to identify the instance class.

IV. RESULTS AND DISCUSSION

In this work, the classification algorithms are applied on various medical datasets and the recognition process is defined. The work is here defined for four dataset obtained from UCI Machine Learning Repository named heart-statlog, diabetes, hepatitis, breast cancer to evaluate the performance of classification techniques using WEKA. Weka expects the datasets to be in ARFF format, so firstly we change the format of these datasets into ARFF format and then test classification techniques on every data set. After the conversion of datasets into ARFF format, their detailed information is given in the table below:

Table 1 : Dataset Description

Sr. No.	Dataset Name	Instances	Attributes
1	heart-statlog.arff	270	14
2	diabetes.arff	768	9
3	breast-cancer.arff	286	10
4	hepatitis.arff	155	20

The work is been applied on four different datasets so that the effective classification process will be performed. The results obtained from the work based on the classification algorithms are given here under

Here figure 2 is showing the results obtained on heart dataset. The figure shows that the bayes algorithm has provided most effective results where as the random tree approach is worst for this dataset.

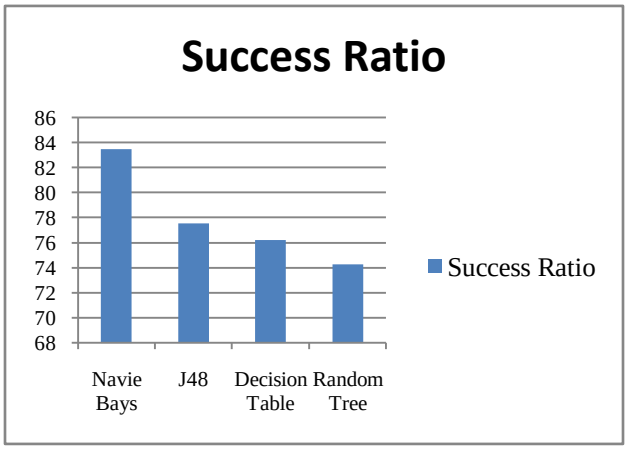

Figure 2 : Recognition Ratio Analysis (Heart Dataset)

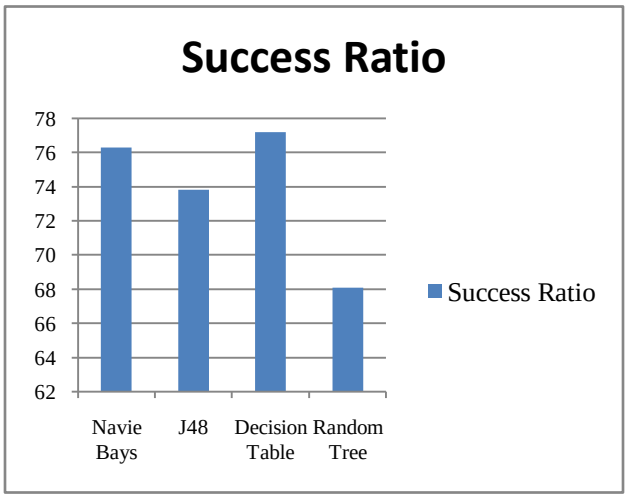

Figure 3 : Recognition Ratio Analysis (Diabetic Dataset)

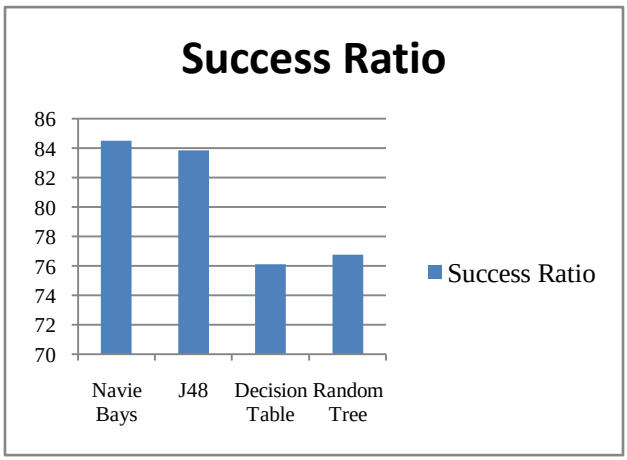

Figure 4 : Recognition Ratio Analysis (Hapatitis Dataset)

Here figure 3 is showing the results obtained on diabetic dataset. The figure shows that the decision tree algorithm has provided most effective results and follwed by bayes algorithm where as the random tree approach is worst for this dataset.

Here figure 4 is showing the results obtained on hapatitis dataset. The figure shows that the bayes algorithm has provided most effective results where as the random tree approach is worst for this dataset.

Here figure 5 is showing the results obtained on breast cancer dataset. The figure shows that the bayes algorithm has provided most effective results where as the decision tree approach is worst for this dataset.

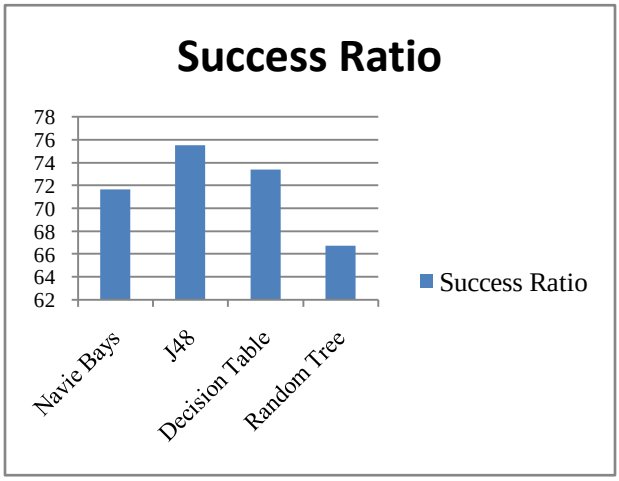

Figure 5 : Recognition Ratio Analysis (breast cancer Dataset)

V. CONCLUSION

The presented work is defined as the analytical study on different algorithmic approaches on medical datasets. Here four different algorithms are analyzed on four different medical datasets. The results shows that the Bayesian tree algorithm provided the most reliable results whereas the results of random tree algorithm are worst.

VI. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami, “Mining Association Rules Between Sets of Items in Large Databases”, In Proceedings of the ACM SIGMOD Conference on Management of data: 207-216, May 1993

[2] Joseph McKendrick, “Big Data, Big Challenges, Big Opportunities: 2012 IOUG Big Data Strategies Survey”, IOUG, Sept 2012

[3] R. Srikant, R. Agrawal, “Mining Quantitative Association Rules in Large Relational Tables”, The 1996 ACM SIGMOD International Conference on Management of Data, Monreal, Canada, pp-1-12, June, 1996

[4] “Big Data for Development: Challenges and Opportunities”, Global Pulse, May 2012

[5] Nigel Wallis, “Big Data in Canada: Challenging Complacency for Competitive Advantage”, IDC, Dec 2012

[6] Ivanka Valova, Monique Noirhomme, “Processing Of Large Data Sets: Evolution, Opportunities And Challenges”, Proceedings of PCaPAC08

[7] Neha Saxena, Niket Bhargava, Urmila Mahor, Nitin Dixit, “An Efficient Technique on Cluster Based Master Slave Architecture Design”, Fourth International Conference on Computational Intelligence and Communication Networks, 2012

[8] Edmon Begoli, James Horey, “Design Principles for Effective Knowledge Discovery from Big Data”, Joint Working Conference on Software Architecture & 6th European Conference on Software Architecture, 2012

[9] Kapil Bakshi, “Considerations for Big Data: Architecture and Approach”, IEEE, 2012

[10] Lawrence 0. Hall, Nitesh Chawla , Kevin W. Bowyer, “Decision Tree Learning on Very Large Data Sets”, IEEE, Oct 1998

[11] Jesmin Nahar," Significant Cancer Risk Factor Extraction: An Association Rule Discovery Approach", Proceedings of International Workshop on Data Mining and Artificial Intelligence (DMAI’ 08) 1-4244-2136-7/08 ©2008 IEEE

[12] Carlos Ordonez," Mining Constrained Association Rules to Predict Heart Disease", 0-7695-1119-8@ 2001 IEEE

[13] Adepele Olukunle," A Fast Algorithm for Mining Association Rules in Medical Image Data", Proceedings of the 2002 IEEE Canadian Conference on Electrical & Computer Engineering 0-7803-7514-9/02@ 2002 IEEE