Comparative Study of Classification Algorithm on Heart Disease Dataset

[1]Janhavi C. Deshpande , [2]Dr. prof. M. A. Pradhan, [3]Sonal S. Hadawale
[1]Student, [2]Senior Professor, [3]Student
[1]Computer Engineering,
[1] All India Shri Shivaji Memorial Society
College of Engineering, Pune-01, India

Abstract : Data classification is process of dividing the dataset into two or more different classes where each class contains similar type of data items. In this paper, we compare different classification algorithms using WEKA tool. Our goal is to analyze the performance of several classifiers on Heart Disease dataset. The analysis is done using six different classification algorithm. From the results of classifiers we found that KNN and J48 are the effective classifiers for medical dataset than other classifiers we have used. Weka provides implementations of learning algorithms that you can easily apply to our dataset. It also includes a variety of tools for transforming datasets, such as the algorithms for classification. You can preprocess a dataset, feed it into a learning scheme, and analyze the resulting classifier and performance.

IndexTerms **- KNN, Decisiontree, Naïve Bayes, DTNB, Spegasos, Classifier**

I. INTRODUCTION

Now a day different soft computing technique is widely used in medical diagnosis. The problem is medical science is in evolution in correct diagnosis as per available information form of data taken from patients. But some soft computing methods are intelligence system and are helpful for classification. For better diagnosis of diseases so many test are needed, these test required classification of large scale data. Data classification is a method of dividing data set into two or more different classes according to the data sets and data features. Features can be selected depends on datasets or application. There are so many classification algorithms in WEKA like Decision tree (J48), KNN (IBK), Naïve Bayes, Adaboast, DTNB, Spegasos etc. The dataset is first trained and then tested the data for classification .The trained data is provided as input to classification algorithm for learning the classifier and hence result are stored. For learning classifier testing data gives as an input for classification.

.

II. CLASSIFICATION ALGORITHM

Classification is a process of data analysis used for extracting a model for learning and making the classes of given data objects,based on that prediction will be made for oblects whose class label is unknown. The classification is done in main two steps training data and testing data. Classification model also represented using mathematical modeling KNN,Decision tree(J48) etc.Some methods of classification are described below:

2.1 Naive Bayes :

1. Naive Bayes(NB) is bayes therom based probabilistic classifier.
2. Naive Bayes classifier produces probability estimates and not the predictions.
3. For each class value they estimate the probability that a given instance belongs to that class. [1]
4. It requires small amout of training data for classification and it is an advantage of this classifier.
5. In this classifer effect of one attribute value is not dependent of the values of another attributes.It is called as conditional independence.

Product rule $P(A \wedge B) = P(A \mid B)P(B) = P(B \mid A)P(A)$

Sum rule $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

Bayes theorem $P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$

Theorem of total probability, if event Ai is mutually exclusive and probability sum to 1.

$$P(B) = \sum_{i}^{n} P(B \mid A_i)P(A_i)$$

Fig 2.1: Naïve bayes classification

Given a hypothesis *h* and data *D* which bears on the hypothesis:
P(h): independent probability of *h*: *prior probability*
P(D): independent probability of *D*
P(D|h): conditional probability of *D* given h: *likelihood*
P(h|D): conditional probability of *h* given *D*: *posterior probability*

2.2 Spegasos:

Spegasos is a technique of Shalev-Shwartzetal[2]. It uses and implements the stochastic variation.In this technique missing values get replaced and nominal attributes are transformed into binary. All attributes are normalized with the output coefficients are based on the values[2].

2.3 IBK or K-Nearest Neighbor (KNN):

k-NN classifer classifies the instances on the basis of similarity. It is popular algorithm for recognition of patterns. k is always considered as positive integer and by majority of the neighbors objects are classified. k-NN is lazy type of learning and in weka it is called as IBK. k-NN algorithm works robustly for noisy data by performing averaging[1]. The k-Nearest Neighbor classification algorithm is based on he Euclidean distance of a test sample and the given training samples,The Euclidean distance between p and q is the length of the line segment connecting them. In cartesian coordinates,if p=(p1,p2...pn) and q =(q1,q2...qn) are two points in euclidean n-space then the distance from p to q or from q to p is given by following:

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Fig 2.3: K-NN Formula

2.4 J48:

J48 algorithm uses divide-and-conquer strategy for creation of decision tree using greedy algorithm.it creats tree with the help of recursion.The tree consist of consists of the root node, branches, parent nodes, child nodes and leaf nodes. [7]. node in a tree denotes dataset attributes; every child node derives labeled branches about the possibilities of attribute values from the corresponding node called parent node.[3]

2.5 AdaBoost:

The Adaboost is very suitable for real time application. Adaboost can be formed with fewer features.AdaBoost improve the classification accuracy and also reduces the processing time. The real Adaboost algorithm gives minor error rates than the diverse AdaBoost.[4]

Adaboost classifier is a combination of weak classifiers and used to construct a strong classifier:

$$f(x) = \sum_{t=1}^{T} \alpha t\, ht(x)$$

where ht:X-> {1,-1} and αt ϵ R

Fig. 2.5: Adaboost

A weak classifier is a very simple model that has just slightly better accuracy than a random classifier, which has 50% accuracy on the training data set. The set of weak classifiers is built iteratively from the training data over hundreds or thousands of iterations.

2.6 DTNB :

This is for building and using a decision table/naive bayes hybrid classifier. At each point in the search, the algorithm evaluates the merit of dividing the attributes into two disjoint subsets: one for the decision table, the other for naive Bayes. A forward selection search is used, where at each step, selected attributes are modeled by naive Bayes and the remainder by the decision table and all attributes are modeled by the decision table initially [5]. At each step, the algorithm also considers dropping an attribute entirely from the model.

III. HEART DISEASE DATASET

For this experiment one medical dataset are taken into consideration, which are collected from UCI repository in ARFF format.Datasets are sufficient for classification process.Datasets are analyzed under different classification technique and parameters.All detail information about dataset like instances and attributesin given table.

No.	DataSet	Instance	Attributes
1	Heart-statlog.arff	270	14

Table 1.Total Instances

3.1 Dataset Description:

Table3.1 : Dataset Information

Dataset of STATLOG are collected from UCI repository. There are 14 attributes in which the last attribute is the class. Remaining 13 attributes namely age, sex, cp, restbps, cholestrol, and blood sugar in fasting ,resting ECG value, Thalach-maximum heart rate achieved, Exang-exercise induced Angina, Old peak value, Slope value, ca, and thal. The Statlog dataset has a prediction attribute with two classes 1 and 2 where 1 represents absence of heart disease and 2 represents presense of heart disease.

IV. ACCURACY

Accuracy of classification algorithm is nothing but the number of correctly classified instances.
Table no.4 shows the accuracy of different classification algorithm Accuracy of classifier is given by:
Accuracy = Correctly Classified Instance / Total number of Instance
Higher the accuracy classifier is more effective:

Classification Algorithm						
Dataset	Naïve Bayes	Spegasos	IBK	Adaboost	DTNB	J48
Heart-Statlog.arff	84.07	85.18	75.92	82.22	78.14	78.88

Table 4. Accuracy of Classification Algorithm

V. RESULTS AND DISCUSSION:

Number of Classified Instances accuracy consisting of number of correctly classified and incorrectly classified instances by six classification algorithms using WEKA tool.

Classifiers	Without Processing	With Processing	Attribute selection	Without Attribute Selection
Naïve Bayes	83.70	83.33	84.07	83.70
Spegasos	82.96	84.81	85.18	83.33
IBK	75.18	75.18	75.92	72.96
Adaboost	80	84.44	82.22	84.44
DTNB	81.48	81.48	78.14	81.11
J48	76.66	79.66	78.88	77.03

Table 5. Accuracy

VI. CONCLUSION

In this paper,the analysis performance of different classification algorithms such as KNN, J48, DTNB, Adaboost, Spegoasos based on accuracy , feature selection and preprocessing. The result we got in analysis shows that KNN and J48 are the best classification algorithms.

REFERENCES

[1] Data Mining Concepts and Techniques- 2nd Edition by Jiawei Han and Micheline Kamber

[2] Shwarz SS, Singer Y, Srebro N. Pegasos: Piramal estimated sub-gradient solver for SVM. In: Proceedings of the 24th International conference on machine learning; 2007. p. 807-814.

[3] Kim G, Lee S, Kim S. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications 2014; 41(4): 1690–1700.

[4] Er.Ramanpreet Kaur, Dr. Vinay Chopra" Implementing Adaboost and Enhanced Adaboost Algorithm in Web Mining" International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 7, July 2015

[5] Devasena, C. Lakshmi, et al. "Effectiveness Evaluation of Rule Based Classifiers for the Classification of Iris Data Set." Bonfring International Journal of Man Machine Interface 1.Special Issue Inaugural Special Issue (2011): 05-09

[6] Mangesh Metkari,M.A.Pradhan"Comparative Study of Soft Computing Techniques on Medical Datasets"Internatioanl Journal of Science and Research(IJSR),ISSN(online):2319-7064,(2012)

[7] Ghorbani AA, Lu W, Tavallaee M. Network Intrusion Detection and Prevention Concepts and Techniques, Springer New York Dordrecht
Heidelberg London; 2010.