Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 02, February -2018

AN EFFICIENT COMPARISION OF DATA CLASSIFICATION ALGORITHM FOR ANALYSIS OF IRIS DATA SETS

[1]R.S.SANJUVIGASINI, [2]DR.R.SHANMUGAVADIVU

[1]Department of Computer Science, PSG College of Arts and Science,Coimbatore,India
[2] Department of Computer Science, PSG College of Arts and Science,Coimbatore,India

Abstract- *Data Mining techniques are helpful in finding out patterns between data attributes and the results in probalistic prediction of the label attribute.Classification is the major task in data mining. In this paper we discuss about comparing the Decision Tree and Naïve Bayes classification algorithms. The Example data set used from repository sited depending upon the number of instance. We allpy it on different data set to analysis of accuracy of the algorithms. This paper helps to get a clear idea on this algorithm which is based on the evaluation of various methodology driven by Rapid Miner tool while equating Precision, Recall and Accuracy.*

Keywords- *IRIS Data Sets, Naïve Bayes, Decision Tree, The RapidMiner tool*

I. INTRODUCTION

Classification of a label attribute is choosen based on the number of instance. Each of the datasets is categorised on IRIS datasets. We classify this iris datasets with the two different classifier Naïve Bayes and Decision Tree.This process helps in finding patterns between and results in a probalistics predictive attribute. The parameters for judging the algorithms are precision, accuracy and recall. It is helpful when training data used instead of testing the data. To find the value of accuracy, precision , recall of the particular algorithm. For study purpose we implement iris dataset and compare precision ,recall Roc Curve parameters.

Data Mining is the field of finding knowledge for identifying and classifying several data. Classification is the process of grouping or spliting data into similar groups based on the criteria. By using this classification techniques we are finding the accuracy of data which we gathered. Classification in data mining is the task to identify the accuracy, and predict to new classes using several techniques. By using Data mining tool we compare the two classification algorithm for our convinent.

Classification of large database can be little difficult but by using Naïve Bayes and Decision Tree Induction algorithm to classify IRIS data based on the accuracy of data. By comparing this two classification algorithms we simply identify the performance, accuracy of each algorithm which finally show the better algorithm.

The main objective is to find the accuracy of data based on the data set which we gave. Classification is has become more important in finding the accuracy of large data sets.The IRIS data we use to compare the attribute and to measure the length and width such as Sepals and Petals. This is named from the flower IRIS. The attributes where considered as the petels of this iris flower.

Whereas analysis of these data sets using data mining techniques are used to find and predict the data Accuracy. Which in turn helps us to preserve our data safe and secure.Different range of data sets may results different accuracy depends on tools used for implementation.Eventhough applying classification techniques to assist cloud data leads to the privacy concern into to the great demand. Here we define accuracy of data and also the performance calculation of two algorithms.

Here we are using the Rapid Miner software to predict the accuracy of Classification algorithms. In this comparsion we use Lift Chart and ROC curve in which it displays the Accurate value of classification theorm.

II. LITERATURE REVIEW

Author Li Liu, Murat Kantarcioglu and Bhavani Thurasingham discussed about the securing of data using decision tree algorithm. . It is classified with the perturbed data set, and this process improves the accuracy of data. It also reduce the costs off communicatio and computation compared to any other cryptographici services They also provide the direction for mapping the data mining functions instead of reconstructing the original data which provide more privacy with less cost [3].

Author Ahmad Ashari, Paryudi, Min tjoa describes about the performance of various classification algorithm for an alternative design in an energy simulation tool. This shows there is possible way of comparing multiple algorithms. As per the comparision of decision tree, naive bayes, K-Nearest Neighbour algorithm the accuracy of decision tree is better than the other algorithms [4].

Author Sagar S.Nikam has defined the comparitive study on classification techniques which mainly focus in the performance analysis of classification algorithms and its Limitations. Also focus on classifying data into different classes according to some constraint. The first approach is the Statistical approach which is classical approach works on linear discrimination. The second is Machine Learning which helps to solve more complex problems and third approach is Neural Network shows the diverse source ranging from the understanding and emulating the human brain to border issues of human abilities [6].

Author Rachna Raghuwanshi has describe about performance of the Naïve bayes classifier and Decision Tree with the Fire Data Set to compare the accuracy. Where as the problem with Cross Validation is avoided [7].

Author XHEMALI, J.HINDE, G.STONE precises on the automatic analysis and classification of attribute data from training course web pages. They choose Naive bayes, Decision Tree, Neural Network algorithm to classify the best data with same data set. As per the result gained the accuracy of naive bayes is more accurate than any other classification algorithm [8].

Author Bhaskar N.Patel, Satish G. Prajapati, Dr.Kamaljit I. Lakhtaria describes the classification is the categorization of data into different category based on some rules. The classification of data with decision tree is the pictorial view, and categorizing is easier, accuracy is better than othe classification algorithm [11].

III. METHODOLOGY

3.1 INTRODUCTION

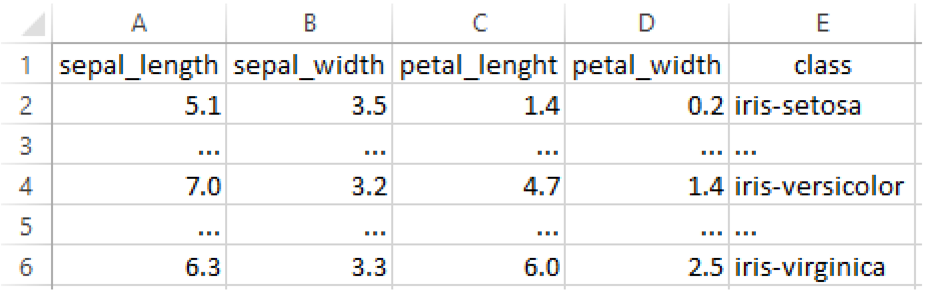
Classification of one label attribute is choosen based on the number of instance. Each of the data sets is IRIS dataset. The IRIS data set is classified under the tool called Rapid Miner. This process helps in finding out the patterns between and results in a probalistics predictive attribute. The two different datasets are used from the repository sites based on the number of instances. These instances where applied in a two different classifiers like Decision Tree, Naive Bayes to identify the Precission, Accuracy, Recall for large datsets.

3.1.1 IRIS DATASETS

IRIS DataSet is the multivariate data set which was introduced by British statistician and biologist RonaldFisher. It is also known as Edgar Anderson collected the data to quantify the mophologic variation of Iris Flower of three related specises.

Figure 3.1 IRIS FLOWER

The file *iris.csv* contains the data for this example in comma sepparated values (CSV) format. A sample of the contents of that file is listed below.



	A	B	C	D	E
1	sepal_length	sepal_width	petal_lenght	petal_width	class
2	5.1	3.5	1.4	0.2	iris-setosa
3	...	...	...	...	...
4	7.0	3.2	4.7	1.4	iris-versicolor
5	...	...	...	...	...
6	6.3	3.3	6.0	2.5	iris-virginica

Figure 4.2 Iris flowers dataset.

The arrtibute value of IRIS data are:

1. **sepal_length**: Sepal length, in centimeters, used as input.
2. **sepal_width**: Sepal width, in centimeters, used as input.
3. **petal_length**: Petal length, in centimeters, used as input.
4. **petal_width**: Petal width, in centimeters, used as input.
5. **setosa**: Iris setosa, true or false, used as target.
6. **versicolour**: Iris versicolour, true or false, used as target.
7. **virginica**: Iris virginica, true or false, used as target.
8.

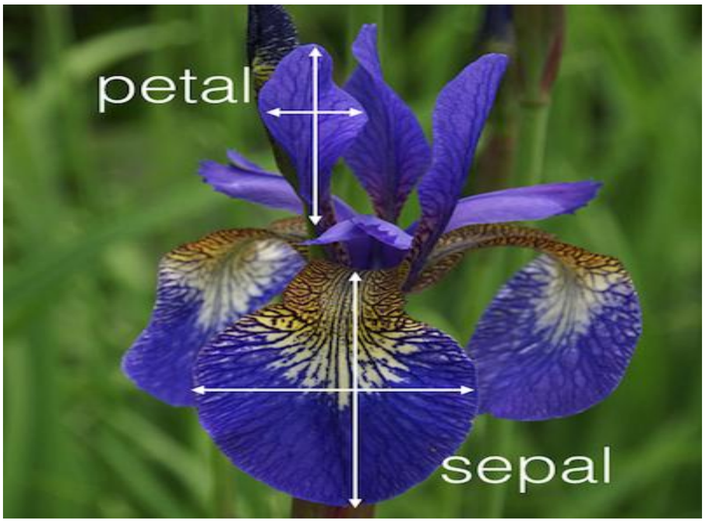


Figure 3.2: IRIS Flower Sepal (Length and Width) and Petal (Length and Width)

3.2 Naïve Bayes Classifier

NB Classifier is used as a knowledge accumulator during training and testing data. This helps to classify and sense unseen data. It also identify and responsible for extracting all suitable features. NB classifiier calculate the most possible output based on input. Naïve bayes is the bettrt probabilistic classifier it consider the presence of a particular features of a class.

1. Let D be the training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimentional attribute vector, X=(x_1, x_2, x_3,….x_n) depictiong n measurements made on the tuple from n attributes, respectively, A_1,A_2,A_3,….A_n.
2. If there are m classes, C_1,C_2,C_3,…..$C_{m.}$ Given tuple ,X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the Naïve Bayes classifier predicts that tuple X belongs to the class C_i if and only if

   $P(C_i|X)> P(C_j|X)$ for $1 \leq j \leq m, j \neq i$.

3.3 Factor considered for calculating Performance of Classifiers

Accuracy of classifiers are compared based on the Precision, Accuracy, Recall, False Positive rate ande True Negative rate. The RapidMiner tool provide powerful platform which gives integrated environment for data mining. The average measure is taken as the overall maesure for classifiers.

The over all precision for a classifier for a given dataset, average of precision of both classes is calculated. Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naive Bayes classifier considers that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) * P(x_2|c) * \ldots\ldots\ldots *P(x_n|c)*P(c)$$

P(c|x) is the posterior probability of class given predictor od class.
P(c) is called the prior probability of class
P(x|c) is the likelihood which is the probability of predictor of given class
P(x) is the prior probability of predictor of class.

3.3.1 Accuracy

Accuracy is the calculation of number of instance predicted positively divided by Total number of Instances.That is accuracy is the percentage of the accurately predicted classes among the total classes.The accuracy is defined as Accuracy = ((True Positive + True Negative)/ (P + N))*100

3.3.2 Precision

Precision is the excat value of true class x which is known as positive predictive value.The proportion of having true positive and the total classified as class x. Precision = (True Positive.(True Positive + False Positive)) * 100

3.3.3 Recall

Recall deals with the sensitive data.It returns the most relevant data and the part of document which is relevant as the result from query. Recall = (True Positive.(True Positive + False Negative)) * 100

3.3.4 True Positive

True positive are the positive tuples which are correctly labelled by the classifiers. Proportion categorized as class x.Projected by the module that are Predicted positively as resultsTrue Positive rate = (True Positive/(TruePositive + False Negative))*100

3.3.5 False Positive

False Positive is th proportion incorrectly cateforized as class x or the actual total of classes, except X.It is incorrectly predicted compared original results.False Positive rate = (False Positive/ (False Positive+True Negative))*100

3.3.6 F-Measure

It is categorized to for F-measure by combining Precision and Recall.

3.4 Performance Measure

The experiments in this research are evaluated using the standard metrics of accuracy, precision, recall and fmeasure for Web Classification [8]. These were calculated using the predictive classification table, known as Confusion Matrix (Table 4.1).

		PREDICTED	
		IRRELEVANT	RELEVANT
ACTUAL	IRRELEVANT	TN	FP
	RELEVANT	FN	TP

Table 3.1: Confusion Matrix

Considering Table 3.1:

TN (True Negative) → Number of correct predictions that an instance is *irrelevant*
FP (False Positive) → Number of incorrect predictions that an instance is *relevant*
FN (False Negative) → Number of incorrect predictions that an instance is *irrelevant*
TP (True Positive) → Number of correct predictions that an instance is *relevant*

IV. RESULTS AND DISCUSSION

Implementation of Comparision of Decision Tree and Naive Bayes algorithm using Rapid Miner. All the classifiers were trained and tested and consisting of a total of abouve 4000 unique features. The naïve bayes classifier gives the highest accuracy of **95.20%** where as Decision Tree gives **98.9%** accuracy**.** As per our research we have proved that the data which we choose is more accurate than Naïve bayes algorithm. In which theLift and ROC chart shows that the data is more secured while comparing to Naïve bayes algorithm. Because the tree model is more easy and accurate to calculate the data.

REFERENCES

[1] Jiawei Han, Micheline Kamber "Data Mining Concepts and Techniques" in proceeding of second edition Morgan Kaufmann Publisher An imprint of Elsevier 2006.

[2] Alka Gangrade, Ravindra Patel " SMC Protocol for Naïve Bayes classification over Grid Partitioned Data using Multiple UTPs" International Journal of Computer Applications(0975 – 8887) Volume 64- No 6. February 2013.

[3] Li Liu, Murat Kantarcioglu and Bhavani Thurasingham"A Novel PrivacyPreserving Decision Tree Algorithm" Technical Report October 2006.

[4] Ahmad Ashari Iman Paryudi A Min Tjoa "Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool"(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No. 11, 2013.

[5] Ashmeet Singh, R Sathyaraj "A Comparison Between Classification Algorithms on Different Datasets Methodologies using Rapidminer" International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 5, May 2016.

[6] Sagar S. Nikam "A Comparative Study of Classification Techniques in Data Mining Algorithms" International Conference on Computer Science and Electronics Engineering 2012

[7] Rachna Raghuwanshi"A Comparative Study of Classification Techniques for Fire Data Set" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (1), 2016, 78-82

[8] Daniela XHEMALI, Christopher J. HINDE and Roger G. STONE**"** Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages" IJCSI International Journal of Computer Science Issues, Vol. 4, No. 1, 2009 ISSN (Online): 1694-0784 ISSN (Print): 1694-0814

[9] Josip Mesarić, Dario Šebalj"Decision trees for predicting the academic success of Students" Croatian Operational Research Review CRORR 7(2016), 367–388 December 30, 2016.

[10] N Raveendran, Dr Antony Selvadoss Dhanamani "Impact of Cloud Computing on Data Mining System" International Journal of Advanced Research in Computer Science Volume 3, No. 6, Nov. 2012 (Special Issue), ISSN No. 0976-5697.

[11] Bhaskar N . Patel, Satish G Prajapati and Dr. Kamaljit I. Lakhtaria" Efficient Classification of Data Using Decision Tree" International Journal of Data Mining, Vol. 2, No. 1, March 2012.

[12]Bharat BhargavaAnya Kim, YounSun Cho "Research in Cloud Security and Privacy"
"https://www.cse.unr.edu/~mgunes/cpe401/cpe401sp14/12-cloud-security.ppt".

[13] S.L. Ting, W.H. IP, Albert H.C. Tsang "Is Naïve Bayes a Good Classifier for Document Classification?", International Journal of Software Engineering and Its Applications, Vol. 5, No. 3, July, 2011.

[14] Shahrukh Teli, Prashasti Kanikar "A Survey on Decision Tree Based Approaches in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, 2015.