A Comparitive Study On Different Classification Algorithms Using Airline Dataset

Prasad A. Jagdale,
Bachelors in Computers Engineering
Computer Engineering Dept.
Pimpri-Chinchwad College of Engineering
Savitribai Phule Pune University, India
prasadjagdale24@gmail.com

Deepa Abin
M.E.(Computer Engineering)
Computer Engineering Dept.
Pimpri-Chinchwad College of Engineering
Savitribai Phule Pune University, India
deepaabin@gmail.com

Dr. K. Rajeswari
B.E, MTech, PhD.
Head Of Department (Computer Engineering)
Pimpri-Chinchwad College of Engineering
Savitribai Phule Pune University, India
kannan.rajeswari@pccoepune.org

Abstract—The paper presents comparison of five differentclassification algorithms performed on airline dataset to find out best accuracy. Here the used dataset is consist of 250 different type of records of airline information consisting flight timing, cities, days of weeks, delay timing.Classification algorithms are performed on the data set and the best result is given by Random Forest.

Keywords- *Classification Algorithms,J48(Iterative Dichotomiser)*

*****

I. Introduction

Data Mining is process where data is analyzed to gain more knowledge and information about the specific data. It is a technology used with great potential to help business and companies to focus on the most important information of the data that they have to collect to find out their customer's behaviors. Quick methods are applied in order to extracting data patterns, and this is done by going through predefine steps like "selection of data, cleaning of data, integration, transforming the data and pattern extraction". Many methods are used for extraction data like" Classification, Regression, Clustering, Rule generation, Discovering, association Rule…etc. each has its own and different algorithms to attempt to fit a model to the data. Algorithm is a set of rules that must be followed when solving a specific problem (it is a finite sequence of computational steps that transform the given input to an output for a given problem). The problem can be a machine. Classification algorithms/techniques in data mining are capable of processing a large amount of data. It can predict categorical class labels and classifies data based on training set and class labels and hence can be used for classifying newly available data. Thus it can be outlined as an inevitable part of data mining and is gaining more popularity [1]

II. Classification Techniques

Classification algorithms are mainly designed to find out in which group each data instance is related within a given dataset. Using these algorithms data can be classified into different classes according to some constrains. Different kinds of classification algorithms are present such as C4.5, ID3, k-nearest neighbor classifier, Naive Bayes, SVM, ANN, J48, Decision Stamp, Random Tree and Random Forest.[4]

The paper compares different type of classification algorithms J48, Decision Stamp, Random Tree, Random Forest, and Naive Bayes.

The classification algorithms are performed on airline dataset which is consisting of 250 recordssize, Each algorithm is run through the data set as it's observed that random forest is giving is best result when comparison is performed based on accuracy.

A. J48

J48 other name is Iterative Dichotomiser3 and denoted by ID3, this algorithm uses a greedy technique to induce decision trees for classification and uses reduced error pruning. It has ability to work with categorical and the continuous valued attributes. [7]

B. Random Forest

Random Forest is an ensemble algorithm which was modeled from treesalgorithm and Bagging algorithm. This algorithm can potentially improve classificationaccuracy. It also works well with a data set with a vast number of input variables. The Algorithm begins by creating a combination of trees which each will vote for a class. [8]

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

C. Naïve Bayesain

The naive Bayes algorithm is a simple probabilistic classifier that calculates a collection of probabilities by investigating frequency and combination of values in a given data set. The algorithm is based on applying Bayes theorem with the "naive" assumption of independence between every pair of features.

D. Random Tree

The Random tree is a tree which is formed bystochastic process. Theloop-erased random walk is a model for a random simple path with important applications in combinatory and, in physics, quantum field theory. It is intimately connected to the uniform spanning tree, a model for a random tree.[6]

E. Decision Stump

A decision stump is a machine learning model consisting of a one-level decision tree. That is, it is decision tree with one internal node (the root) which is immediately connected to the terminal nodes (its leaves).A decision stump makes a prediction based on the value of just a single input feature.[2]

III. Performance Measures

There are different formulas present to calculate performance measures out of which for this experiment following performance measures are used, following are 5 performance measures are used here to search the best accuracy.[3]

(Accuracy, Error rate, Precision, Recall, F-Measure)

1. Accuracy

Measures the proportion of correctly classified cases.

$$Accuracy, Acc = \frac{TP+TN}{TP+FN+FP+TN}$$

2. Misclassification Rate (MR)

Incorrectly classified instances were measured using misclassification rate as in, [5]

$$MR = \frac{FP+FN}{TP+FN+FP+TN}$$

3. Error rate

Measures the proportion of incorrectly classified cases.

$$Error\ Rate, Err = \frac{FN+FP}{TP+FN+FP+TN}$$

4. *Precision*

Measures the fraction of true positives against all positive results.

$$Precision, p = \frac{TP}{TP+FP}$$

5. Recall

Measures the fraction of positive cases classified as positive.

$$Recall, r = \frac{TP}{TP+FN}$$

6. F- measure

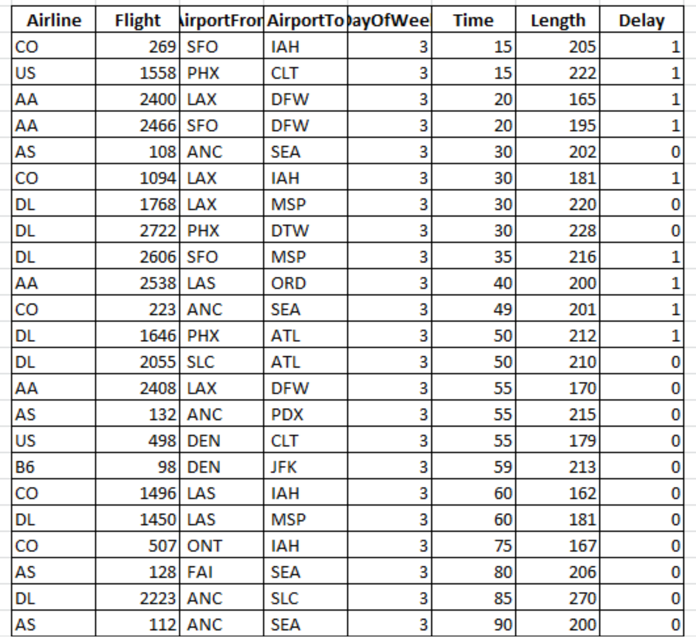
Measures the weighted harmonic mean of its precision and recall.

$$F\text{-}Measures = \frac{2rp}{r+p} = \frac{2TP}{2\ X\ TP+FP+FN}$$

IV. Expriement environment

For Experimentation 5 classification algorithms are selected and airline data set is used as airline consist of flight times, state information, delay timing, flight numbers, and data is related to every other record with most of the time unique records this will create a complex scenarios where finding class major task as result it will be helpful for finding best classification algorithm. To perform this experiment dataset of 250 records from airline database is selected Fig.1.The classification algorithm used in this experiment is as follows J48, Decision Stamp, Random Tree, Random Forest, and Naive Bayes.

Each algorithm is ran through the data set to calculate performance measures, for this experiment following measures are used, Precision, Recall, Accuracy, FP rate, TP rate, F-Measures .



Airline	Flight	irportFror	AirportTo	ayOfWee	Time	Length	Delay
CO	269	SFO	IAH	3	15	205	1
US	1558	PHX	CLT	3	15	222	1
AA	2400	LAX	DFW	3	20	165	1
AA	2466	SFO	DFW	3	20	195	1
AS	108	ANC	SEA	3	30	202	0
CO	1094	LAX	IAH	3	30	181	1
DL	1768	LAX	MSP	3	30	220	0
DL	2722	PHX	DTW	3	30	228	0
DL	2606	SFO	MSP	3	35	216	1
AA	2538	LAS	ORD	3	40	200	1
CO	223	ANC	SEA	3	49	201	1
DL	1646	PHX	ATL	3	50	212	1
DL	2055	SLC	ATL	3	50	210	0
AA	2408	LAX	DFW	3	55	170	0
AS	132	ANC	PDX	3	55	215	0
US	498	DEN	CLT	3	55	179	0
B6	98	DEN	JFK	3	59	213	0
CO	1496	LAS	IAH	3	60	162	0
DL	1450	LAS	MSP	3	60	181	0
CO	507	ONT	IAH	3	75	167	0
AS	128	FAI	SEA	3	80	206	0
DL	2223	ANC	SLC	3	85	270	0
AS	112	ANC	SEA	3	90	200	0

Figure 1. Short Part of Used Airline Dataset

V. Result and Expriement environment

Data set is run through all 5 Classification algorithms J48, Decision Stamp, Random Tree, Random Forest, and Naive Bayes, and performance measures are calculated.

While performing the experimentation it’s been observed that each algorithm gives different result whereas J48 and Naive Bayesian performance measure values are almost close to each other which results in same accuracy.

As 5 classification algorithms are selected for comparison which in results produces 5 set of different performance measures and based on the values from each algorithm graphs are constructed *(Ref. Figure 2,Figure 3,Figure 4,Figure 5 andFigure 6)* which clearly shows that J48 – Naïve Bayes are identical and best performance measure values are given by Random Forest where the lowest performance values are given by decision stamp.

Following are the graphs reflecting the performance measures performed on airline dataset. They are sorted in increasing accuracy wise; as we change the classification algorithms the accuracy got changed and Random Forest gives higher performance measure values followed by Naïve Bayes And J48; the lowest performance measure values are given by Random Tree followed by Decision Stamp.

The Decision Stamp gives 74% accuracy and lowest performance measure values. Fig. 2

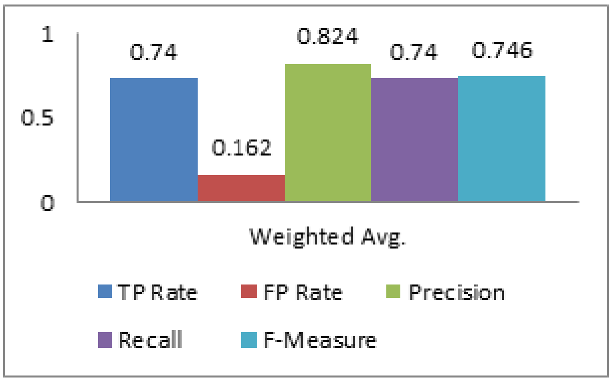


Figure 2. Decision Stamp Performance Measure

The random Tree gives 76% accuracy and 2nd lowest performance measure values.Fig. 3

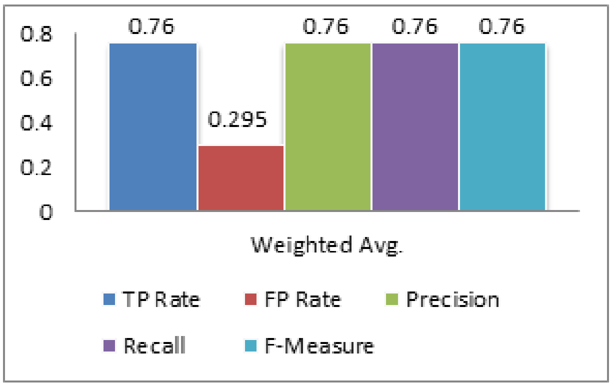


Figure 3. Random Tree Performance Measure

The J48 gives 77% accuracy which is identical to Naive Bayes.Fig. 4

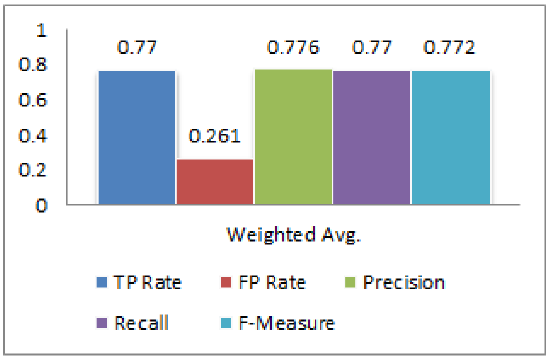


Figure 4. J48 Performance Measure

The Naive Bayes gives 77% accuracy .Fig. 5

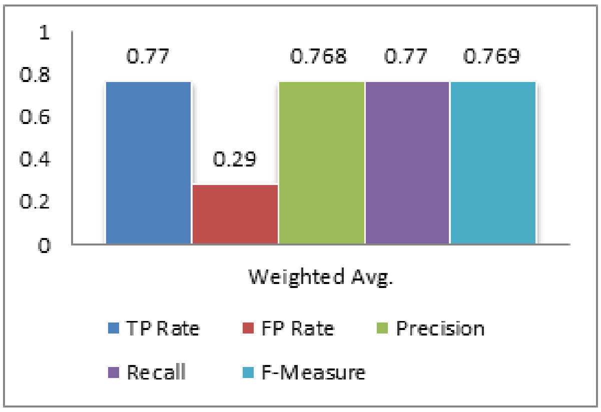


Figure 5. Naïve Bayes Performance Measure

The Random Forest gives 83% accuracy and highest performance measure values.Fig. 6

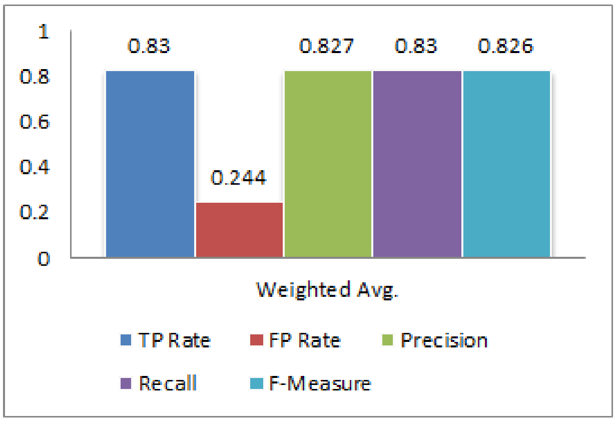


Figure 6. Random Forest Performance Measure

As seen in below chart the graph shows accuracy wise comparison.The lowest accuracy value is given by Decision Stamp which is 74% and highest is given by Random forest which is 83%Fig. 7

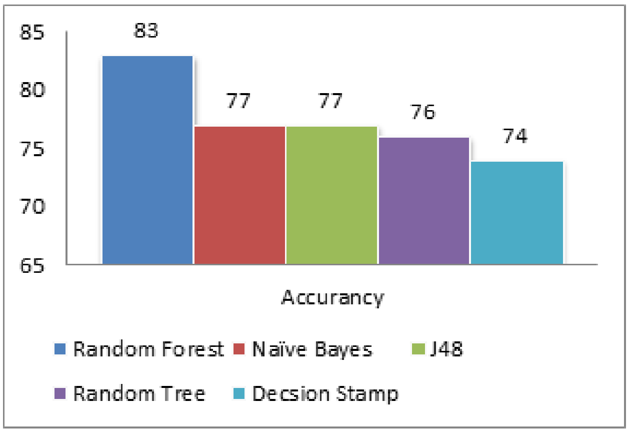


Figure 7.Accuracy Measures Of All 5 Algorithms

VI. CONCLUSION

To classify the data at best accuracy there are many different classification algorithms are present, out of which five algorithms are compared in this paper and computation using all 5 algorithm is performed on the data to find out the best accuracy of algorithm and from this experiment we can conclude that Random Forest gives highest accuracy followed by Naïve Bayes with J48. Random Forest performance measure ishigh as compared with other four classification algorithms, which results in high accuracy.

REFERENCES

[1] Computer Engineering and Intelligent Systems ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)Vol.4, No.8,2013

[2] Iba, Wayne; and Langley, Pat (1992); Induction of One-Level Decision Trees, in ML92: Proceedings of the Ninth International Conference on Machine Learning, Aberdeen, Scotland, 1–3 July 1992, San Francisco, CA: Morgan Kaufmann, pp. 233–240

[3] Jeetesh Kumar Jain, Nirupama Tiwari, Manoj Ramaiya / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 3, Issue 1, January -February 2013, pp.2065-2069

[4] A Comparative Study of Classification Techniques in Data Mining Algorithms Sagar S. Nikam, Department of Computer Science, K.K.Wagh College of Agriculture, Nashik, India.

[5] Feature Extraction with Ordered Mean Values for Content Based ImageClassification , Volume 2014 (2014), Article ID 454876Sudeep Thepade,1 Rik Das,2 and Saurav Ghosh, Pimpri Chinchwad College of Engineering, Akurdi, Sec. 26, Pradhikaran, Nigdi, Pune, Maharashtra 411033, India

[6] Richard Kenyon, The asymptotic determinant of the discrete Laplacian, Acta Math. 185:2 (2000), 239-286, online version.

[7] For an implementation of ID3v1 in Python, see Dive Into Python, Chapter 5. Objects and Object-Orientation

[8] Gareth James; Daniela Witten; Trevor Hastie; Robert Tibshirani (2013). An Introduction to Statistical Learning. Springer. pp. 316–321.