ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 9, Issue, 04, pp.48691-48693, April, 2017

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

PERFORMANCE EVALUATION OF CLASSIFICATION AND CLUSTERING ALGORITHMS ON DATASETS

Dr. Jyotsna Sinha and *Deepika Sharma

R.C Institute of Technology, (Affiliated to GGS IP University, Najafgarh, New Delhi), India

ARTICLE INFO

Article History:
Received 14th January, 2017
Received in revised form
15th February, 2017
Accepted 22nd March, 2017
Published online 20th April, 2017

Key words:

Algorithms, parameters,
Cluster distribution, ROC area ,
f -measure , Kappa Statics.

ABSTRACT

This research paper is a comparative study of different clustering & classification algorithms. Clustering algorithms are compared on the basis of accuracy parameter, cluster distribution and time taken to build model. The six accuracy parameters for evaluating accuracy of classification algorithms are used. These parameters are TP rate, precision, recall, ROC area, f-measure and kappa statics. The four error measurement parameters for evaluating error rate of classification algorithms are considered i) RMSE (Root mean squared error) ii) MAE (Mean absolute error) iii) RRSE (Root relative squared error) iv) RAE(Relative absolute error).

Citation: Dr. Jyotsna Sinha and Deepika Sharma, 2017. "Performance evaluation of classification and clustering algorithms on datasets", *International Journal of Current Research*, 9, (04), 48691-48693.

INTRODUCTION

Data mining is a process to discover interesting knowledge, such as associations, patterns, anomalies, changes and significant structures from large amount of data stored in databases or other information repositories. There are various data mining techniques such as Association, Classification, Clustering, Neural Network and Regression. Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. In classification, we make the software that can learn how to classify the data items into groups. Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristics using automatic technique. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups.

EXPERIMENTAL SETUP

We have compared the performance of both clustering and classification algorithms using weka tool. Weka is a data mining software which contains a set of tools for pre-processing, clustering, regression, classification and visualization of data.

Corresponding author: Deepika Sharma
R.C Institute of Technology, (Affiliated to GGS IP University, Najafgarh, New Delhi), India.

In this work we have compared three clustering algorithms (k-means clustering, DBSCAN, Hierarchical) on the basis of number of clusters, cluster instances, accuracy and time taken to build the model. We have also compared four classification algorithms (J48, OneR, Naïve Bayes, Decision table) on the basis of MAE, RAE, RRSE,and RMSE.

Data Collection

We have used real world data for our work which has been obtained from UCI repository. The clustering algorithms have been applied on bank dataset whereas classification algorithms on diabetes dataset.

Dataset Description

- The sample dataset used for performing clustering is based on "bank data" available in comma-separated version (csv) format. This dataset consists of 12 attributes and 600 instances. A version of data set, i.e. "bank.arff", has been created in which ID field has been removed. Table shown below gives the description of bank dataset.
- The sample dataset used for comparing classification algorithms is "diabetes diagnosis data" available in csv format. This dataset consists of 768 instances and 9 attributes. This dataset has been donated by National Institute of Diabetes and Digestive and Kidney Diseases. In particular, all patients are here female at least 21 years old

of Pima Indian Heritage. There are zeroes which encode missing data for blood pressure attribute. A version of dataset, i.e. "diabetes diagnosis. arff", has been created. Table shown below gives the description of diabetes diagnosis dataset.

Table 1. Bank Dataset Description

Id	a unique identification number
Age	age of customer in years (numeric)
Sex	MALE / FEMALE
Region	inner_city/rural/suburban/town
Income	income of customer (numeric)
Married	is the customer married (YES/NO)
Children	number of children (numeric)
Car	does the customer own a car (YES/NO)
save_acct	does the customer have a saving account (YES/NO)
current_acct	does the customer have a current account (YES/NO)
Mortgage	does the customer have a mortgage (YES/NO)
Pep	did the customer buy a PEP (Personal Equity Plan) after the last mailing (YES/NO)

Table 2. Diabetes Dataset Description

Pregnancies	Number of times pregnant(Numeric)
PG Concentration	Plasma glucose concentration a 2 hours in an oral glucose tolerance test (Numeric)
Diastolic BP	Diastolic blood pressure (mm Hg) (Numeric)
Tri fold thick	Triceps skin fold thickness (mm(Numeric)
Serum Ins	2-Hour serum insulin (mu U/ml(Numeric)
BMI	Body mass index (weight in kg/(height in m)^2) (Numeric)
DP Function	Diabetes pedigree function (Numeric)
Age	Age (years) (Numeric)
Diagnosis	Class variable (0 or 1) (Sick/ Healthy)

RESULTS

Results of bank dataset on different clustering algorithms

Table 3 below shows the experimental results obtained while comparing the clustering algorithms.

Table 3. Clustering algorithms result for Bank Dataset

Algorithm	No. of clusters	Cluster instance	Number of iteration	Time	Accuracy
k-means	2	247(41%) 353(59%)	3	0.08s	56.66%
DBSCAN	2	243(41%) 357(60%)	3	0.03s	56.66%
Hierarchical Clustering	2	599(100%) 1 (0%)	3	0.53s	54.16%

Figures below shows the graphical representation of the results for the comparison of different clustering algorithms on the basis of Accuracy rate, Time taken and cluster distribution.

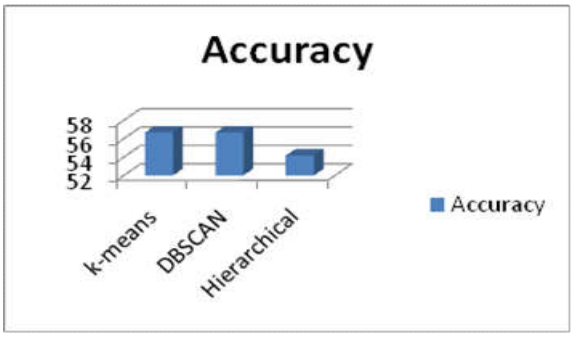

Figure 1. Comparison of Accuracy of K-means, Hierarchical and Density Based clustering

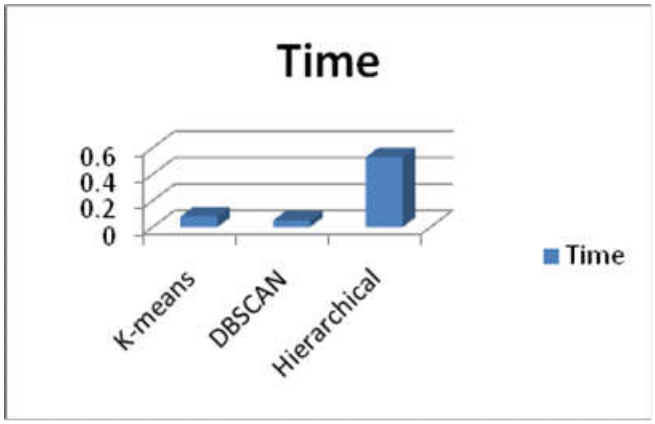

Figure 2. Time take by the K-means, Hierarchical and Density Based clustering for datasets

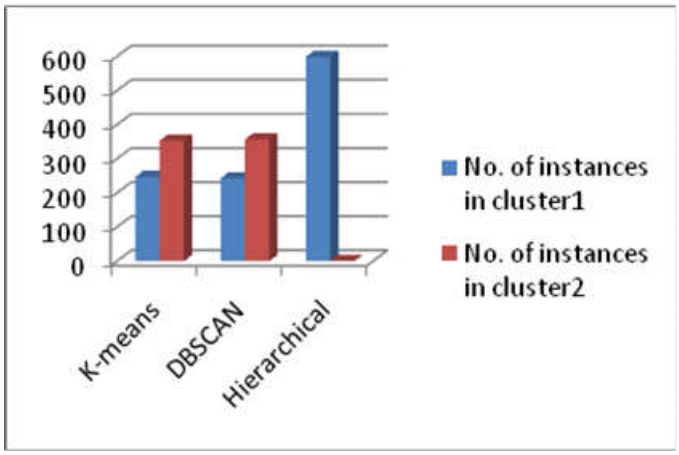

Figure 3. Cluster distribution

Results of diabetes dataset on different classification algorithms

Table 4 shows the six parameters for evaluating accuracy of algorithms. These parameters are Kappa Statics, TP Rate, Precision, Recall, F-measure and ROC area

Table 4. Accuracy Parameters for Diabetes Dataset

Algorithm	Kappa stats.	TP rate	Precision	Recall	F-measure	ROC Area
J48	0.6319	0.841	0.842	0.841	0.836	0.888
Naïve bayes	0.4674	0.763	0.759	0.763	0.760	0.825
Decision Table	0.5432	0.793	0.792	0.793	0.793	0.858
OneR	0.4593	0.764	0.759	0.764	0.758	0.720

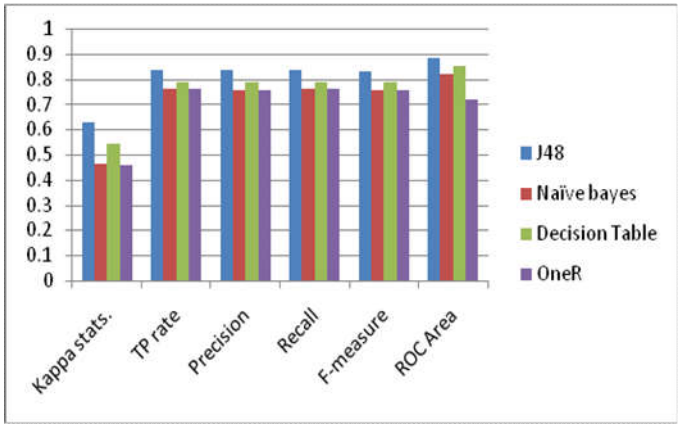

Figure 4. Graphical Representation of Accuracy Parameters

Table 5 shows the four basic error rate parameters for the evaluation of four classification algorithms.

Table 5. Error Rate Evaluation Parameters for Diabetes Dataset

Algorithm	MAE	RMSE	RAE	RRSE
J48	0.2383	0.3452	52.4339%	72.4207%
Naïve bayes	0.2811	0.4133	61.8486%	86.7082%
Decision Table	0.3063	0.38	67.3862%	79.7336%
OneR	0.2357	0.4855	51.8551%	101.8515%

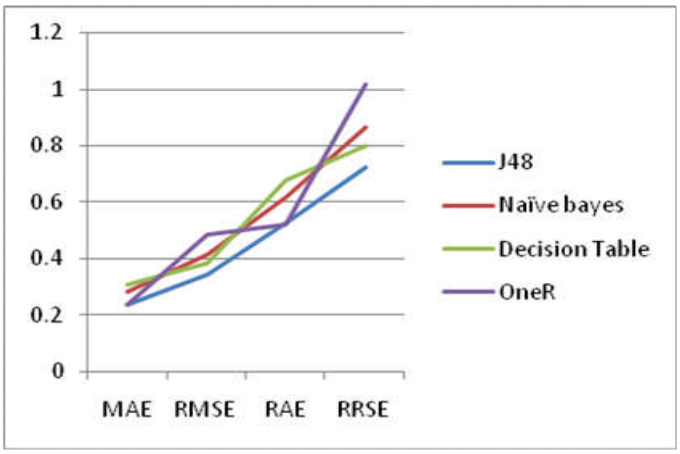

Figure 5. Graphical Representation of Error rate Evaluation

CONCLUSION

Performance based comparative study of clustering and classification algorithms are performed here on two different datasets. The experimental results of various clustering and classification algorithms are depicted in form of tables and graphs. From Figure 1 and Figure 2 it is clear that DBSCAN is the best algorithm as it takes lesser time (0.03seconds) to build the model and gives higher accuracy as compared to other clustering algorithms. It is evident from figure 4 and figure 5 that J48 classification algorithm gives best performance as compared to other studied algorithms. J48 gives higher accuracy rate and minimum error rate. Decision table algorithm has second minimum error rate and it also have over all good performance. As seen in the graph OneR have high error rate and have poor performance as compare to other algorithms under study.

REFERENCES

Abdullah H. Wahbeh et. al. "A Comparison Study between Data Mining Tools over some Classification Methods".International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence, pp. 18-26.

Dr. T.nalini, S. Revathi, 2013. "Performance Comparison of Various Clustering Algorithm". *International Journal of Advanced Research in Computer Science and Software Engineering,* Vol. 3, Issue 2, Feb. pp. 67-72.

Prakash Singh, Aarohi Surya, "Performance analysis of clustering algorithms in data mining in weka". *International Journal of Advances in Engineering & Technology,* Vol. 7, Issue 6, pp. 1866-1873.

Trilok Chand Sharma, Manoj Jain, 2013. "WEKA Approach for Comparative Study of Classification Algorithm". *International Journal of Advanced Research in Computer and Communication Engineering,* Vol. 2, Issue 4, April, pp. 1925-1931.

Vishal Shrivastava, Prem narayan Arya, 2012. "A Study of Various Clustering Algorithms on Retail Sales Data". *International Journal of Computing, Communications and Networking,* Vol. 1. No. 2, September – October, pp. 68-74.

****