IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.6, Special Issue.11, Dec 2018 E-ISSN: 2347-2693

A Comparative Study on Weka, Orange Tool for Mushroom Data Set

M. Senthamilselvi[1*], P.S.S. Akilashri[2]

[1,2]Department of Computer Science, National College Tiruchirappalli, India

Available online at: www.ijcseonline.org

Abstract— Association rule mining (ARM) is used to advance decisions making in the application. ARM became important in an information and decision-overloaded world. They changed the way user make decisions, and helped their creators to increase revenue at the same time. Bringing ARM is essential in order to popularize them beyond the limits of scientific research and high technology entrepreneurship. This paper extracts attractive correlations frequent patterns and association among set of items in the transaction database. This paper describes the show analysis of Naïve Bayes and J48 classification algorithm based on the correct and incorrect instances of data classification. Naive Bayes is probability based and j48 algorithm is decision tree based. In this paper Comparison weka and orange by using tool to perform evaluation of classifiers NAIVE BAYES and J48 in the context of mushroom dataset in UCI repository to maximize true positive rate and minimize false positive rate of defaulters rather than achieving only higher classification accuracy using WEKA and orange tool. The experiments results reveals the true positive rate, false positive rate, classification accuracy and cost analysis. The results in the paper on mushroom dataset in UCI repository performance best in Weka than Orange tools. The efficiency and accuracy of J48 -than Naive Bayes is good.

Keywords: Data mining, Weka Tool, Orange tool J48,Navie Bayes

I. INTRODUCTION

Data mining is a process of extracting interesting knowledge or patterns from large databases. There are several techniques that have been used to discover such kind of knowledge, most of them resulting from machine learning and statistics. The greater part of these approaches focus on the discovery of accurate knowledge. Though this knowledge may be useless if it does not offer some kind of surprisingness to the end user. The tasks performed in the data mining depend on what sort of knowledge someone needs to mine. Data mining technique are the result of a time-consuming process of study and product development. The main types of task performed by DM techniques are Classification, Dependence Modeling, Clustering, Regression, Prediction and Association. Classification task search for the knowledge able to calculate the value of a previously defined goal attribute based on other attributes and is often represented by IF-THEN rules. We can say the Dependence modeling as a generalization of classification. The aim of dependence modeling is to discover rules able to calculate the goal attribute value, from the values of calculated attributes. Even there are more than one goal attribute in dependence modeling. The process of partitioning the item set in a set of significant sub-classes (called clusters) is known as Clustering.

II LITERATURE SURVEY

Zhang et al [1] proposed an algorithm to discover combined association rules. Compared with the existing association rule, this combined association rule technique allows different users to perform actions directly. In this paper they have focused on rule generation and interestingness measures in combined association rule mining. In combine association rule generation, the frequent itemsets among itemset groups are discovered to improve efficiency.

Gautam [2] obtainable an specialized version of Apriori techniques for mining multilevel association rules in large databases to finding maximum frequent itemset at lower level of abstraction. This paper proposed a new, speedy and an efficient algorithm (SCBF Multilevel) with single scan of database for mining complete frequent itemsets. This system algorithm can derive the multiple-level association rules under different supports in simple and effective way.

Bao [3] paper an algorithm for double connective association rule mining for which a three table relational database is used. The rules are found among the primary keys of the two entity tables and the primary key of the binary relationship table.

Yahya Slimani [4] paper a dynamic load balancing strategy for distributed association rule mining algorithms under a Grid computing environment. Experiments showed that the proposed strategy succeeded in achieving better use of the Grid architecture assuming load balancing and this for large sized datasets.

Abdul Kadir et al[5] provide the preliminary of basic concept of negative association rule and proposed an improvement in

Apriori algorithm for mining negative association rule from frequent absence and presence itemset. Relative interestingness measures were adopted to prove that the generated rules are also interesting and strong.

Liu et al [6] existing different method to deal with the false positive errors in association rule mining. Three multiple testing correction approaches- the direct adjustment approach, the holdout approach and the permutation-based approach are used and extensive experiments have been conducted to analyze their performances. From the results obtained, all the three approaches control false positives effectively but among the three permutation–based approach has the highestpower of detect actual association rules, but it is computationally expensive.

Gupta Anekritmongkol [7] paper algorithm algebra algorithm that will decrease the amount of time in reading data from the database. It has been found that through experiments that the time was reduced considerably. Similar authors have compare the performance of unusual association rule mining algorithms by implementing them on various kinds of datasets.

III METHODOLOGY

3.1 Navie Bayes

Naïve Bayes is a supervised probability machine learning classifier method that assumes terms occur independently. This can be used to in classifying textual documents in simplest method. The Naïve Bayes model computes the posterior probability of a class, based on the allocation of words in the document this illustration works with the BOWs feature extraction which ignores the situation of the word in the document .Bayesian classification represent a supervised learning method as well as statistical method for classification. It is easy probabilistic classifier based on Bayesian theorem with strong independence assumption. It is for the most part suited when the dimensionality of input is high. They can predict the probability that a given tuple belongs to a particular class. This classification is named after Thomas Bayes who proposed the bayes theorem. Bayesian formula can be written as

$$\mathbf{P(H \mid E) = [P(E \mid H) * P(H)] / P(E)}$$

The basic idea of Bayes's rule is that the outcome of a hypothesis or an event (H) can be predicted based on some evidences (E) that can be observed from the Bayes's rule.

3.2 J48

J48 is the decision tree based algorithm and it is the extension of C4.5. With this technique a tree is constructed to model the classification process in decision tree the internal nodes of the tree denotes a test on an attribute, branch represent the outcome of the test, leaf node holds a class label and the topmost node is the root node.

```
                Input: C4.5 or J48 Decision Tree T
Procedure PostPruning(Data, TreeSplits)
SplitData (TreeSplits, Data, GrowingSet, PruningSet)
Estimate = DivideAndConquer(Growing Set)
loop
New Estimate = Selection(Estimate,PruningSet)
if Accuracy(New Estimate, Pruning Set) <
Accuracy(Estimate,PruningSet)
exit loop
Estimate = NewEstimate
        return(Estimate)
Procedure Divide And Conquer(Data)
Estimate = Ø
while Positive(Data) != Ø
Leaves = Ø
Instance = Data
while Negative(Instance) != Ø
Leaves = Leaves ∪Find(Leaves, Instance)
Instance = Instance(Leaves, Instance)
Estimate = Estimate ∪ Leaves
Data = Data - Instance
return(Estimate)
```

3.3 WEKA TOOL

Weka is a landmark system in the history of the data mining and machine learning research communities, because it is the only toolkit that has gained such widespread adoption and survived for an extended period of time. The Weka or woodhen (Gallirallus australis) is an endemic bird of New Zealand. It provides many different algorithms for data mining and machine learning. Weka is open source and freely available. It is also platform-independent.

Advantages

- As weka is fully implemented in java programming languages, it is platform independent & portable.
- It is freely available under GNU General Public License.
- Weka s/w contain very graphical user interface, so the system is very easy to access.
- There is very large collection of different data mining algorithms.

Disadvantages

- Lack of possibilities to interface with other software
- Performance is often sacrificed in favour of portability, design transparency, etc.
- Memory limitation, because the data has to be loaded into main memory completely

3.4 ORANGE

Orange Data mining: Orange is an open source data visualization and analysis tool. Orange is developed at the Bioinformatics Laboratory at the Faculty of Computer and

Information Science, University of Ljubljana, Slovenia, along with open source community. Data mining is done through visual programming or Python scripting. The tool has components for machine learning, add-ons for bioinformatics and text mining and it is packed with features for data analytics. Orange is a Python library. Python scripts can run in a terminal window, integrated environments like PyCharm and PythonWin, or shells like iPython. Orange Data mining In Orange, data analysis.

Advantage
Orange tool that are interactive, visual, understandable, well-performing and work directly on the data warehouse/mart of the organization could be used by front line workers for immediate and lasting business benefit.

Disadvantage
The techniques deployed by orange tools are generally well beyond the understanding of the average business analyst or knowledge worker. This is because the tool was generally designed for expert statisticians involved in the detailed science of predictive modeling. This would be the disadvantage today. If this advanced level of analysis is reserved for the few, instead of for the masses, the full value of data mining in the organization cannot be realized. For those with average analytical capabilities, orange is not nearly as effective as it could be.

IV. EXPERIMENT AND RESULTS

The proposed system has been implemented using weka and Orange tool environment. It can be executed on windows. The results are obtained as follows after execution Mushroom data set. Data a mushroom dataset has been used with 119 items each for analysis. A set of association rules are obtained by applying Navie bayes and J48. By analyzing the data and giving different support and confidence values, execution time, can obtain different number of rules. During analysis it found that J48 is much faster for huge number of transactions as compare to Navie Bayes. It takes less time to generate frequent item sets. We work on mushroom data which contains 8124 transactions.

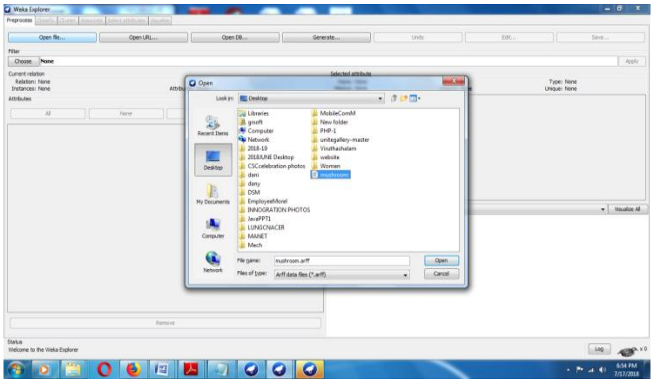
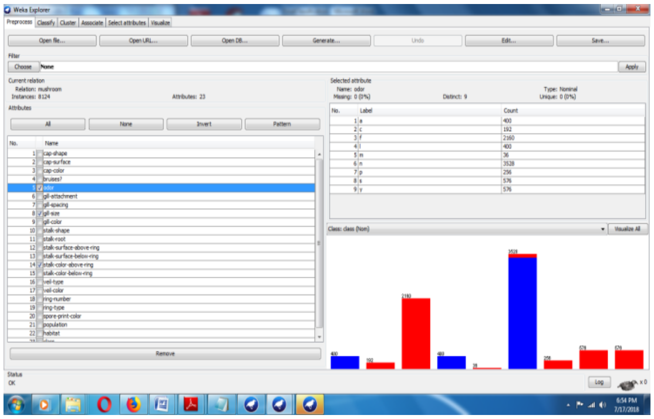


Figure 4.1.1 Upload Mushroom Dataset



No.	Label	Count
1	a	400
2	c	192
3	f	2160
4	l	400
5	m	36
6	n	3528
7	p	256
8	s	576
9	y	576

Figure 4.1.2 View Mushroom Dataset Chart

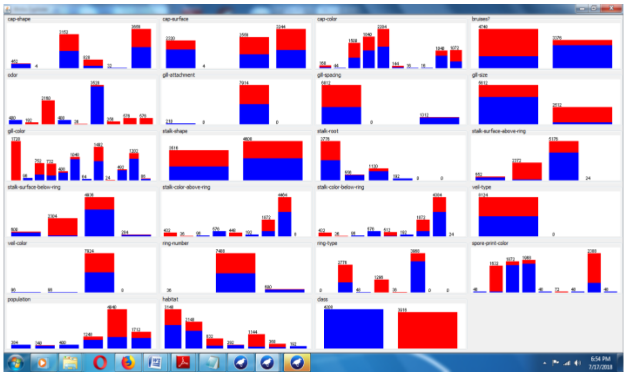
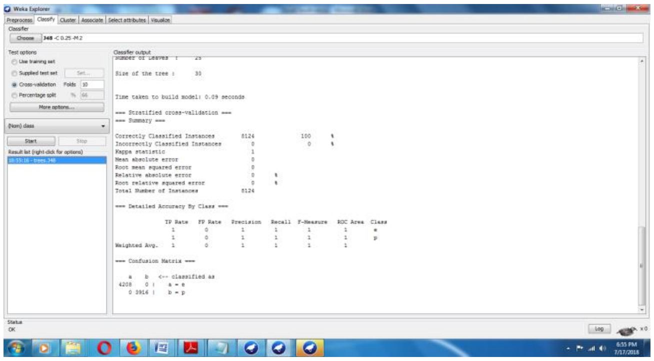


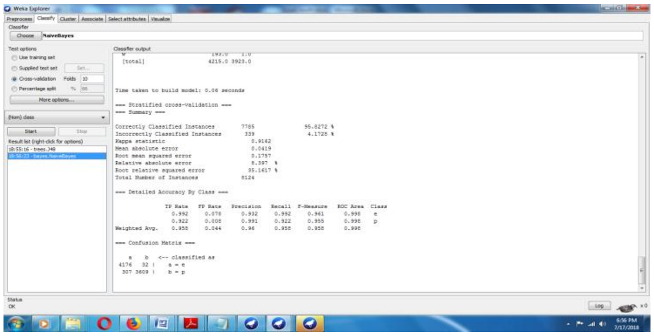
Figure 4.1.3 Over All Chart



J48 -C 0.25 -M 2

```
Number of Leaves  :     25

Size of the tree :      30


Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        8124              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances             8124

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1          1         1          1        e
                 1         0          1          1         1          1        p
Weighted Avg.    1         0          1          1         1          1

=== Confusion Matrix ===

    a    b   <-- classified as
 4208    0 |    a = e
    0 3916 |    b = p
```

Figure 4.1.4 J48 result for weka



NaiveBayes

```
  [total]                   4215.0 3923.0


Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        7785               95.8272 %
Incorrectly Classified Instances       339                4.1728 %
Kappa statistic                          0.9162
Mean absolute error                      0.0419
Root mean squared error                  0.1757
Relative absolute error                  8.397  %
Root relative squared error             35.1617 %
Total Number of Instances             8124

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.992     0.078      0.932     0.992     0.961      0.998    e
                 0.922     0.008      0.991     0.922     0.955      0.998    p
Weighted Avg.    0.958     0.044      0.96      0.958     0.958      0.998

=== Confusion Matrix ===

    a    b   <-- classified as
 4176   32 |    a = e
  307 3609 |    b = p
```

Figure 4.1.5 Navie Bayes result for weka

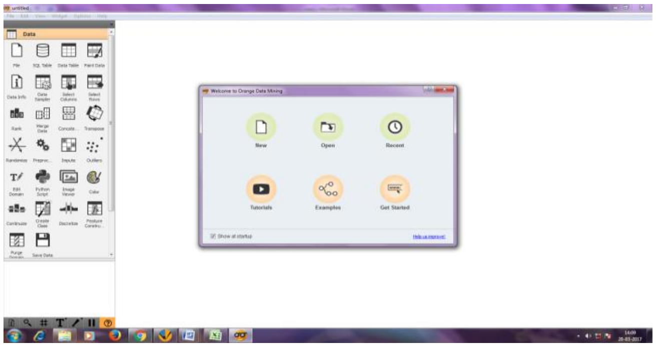
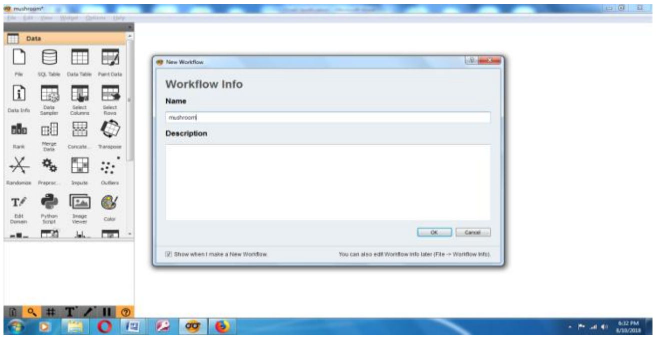


Figure 4.1.6 Orange Tools


New Workflow

Workflow Info

Name

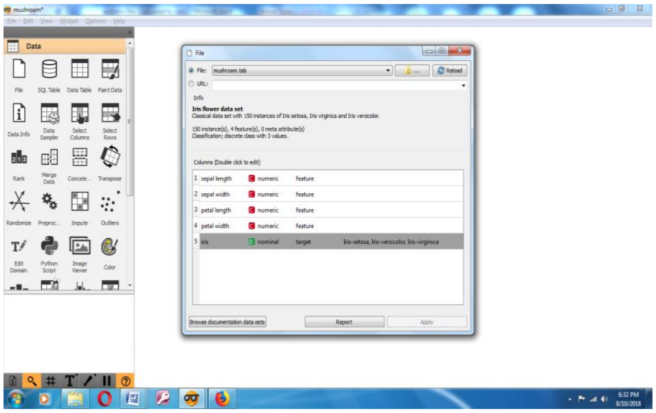
mushroom

Description

☑ Show when I make a New Workflow

You can also edit Workflow info later (File -> Workflow Info).

Figure 4.1.7 Orange Tools Mushroom Dataset


File

● File: mushroom.tab

○ URL:

Info

Iris flower data set
Classical data set with 150 instances of Iris setosa, Iris virginica and Iris versicolor.

150 instance(s), 4 feature(s), 0 meta attribute(s)
Classification; discrete class with 3 values.

Columns (Double click to edit)

1	sepal length	numeric	feature	
2	sepal width	numeric	feature	
3	petal length	numeric	feature	
4	petal width	numeric	feature	
5	iris	nominal	target	Iris-setosa, Iris-versicolor, Iris-virginica

Figure 4.1.8 Orange Tools Mushroom Dataset View

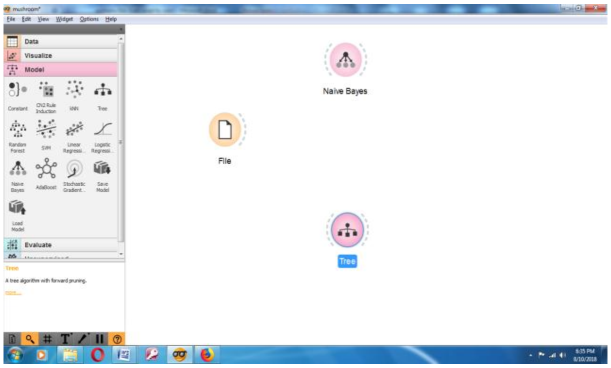


Figure 4.1.9 Algorithm import orange tool

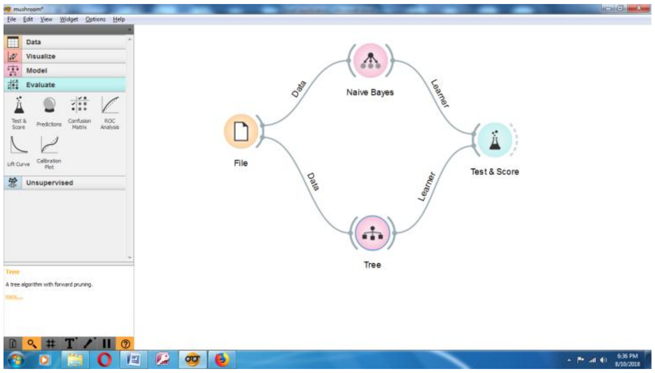
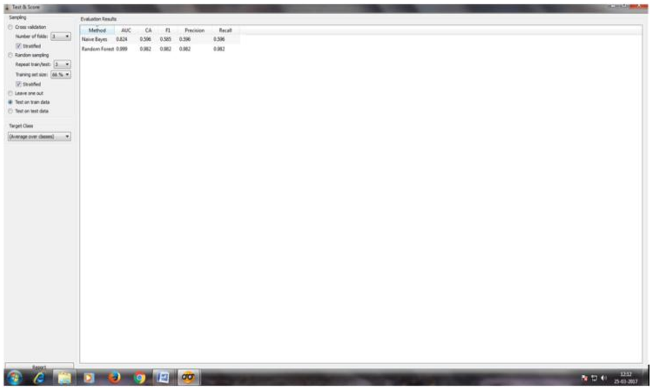


Figure 4.1.10 Algorithm import orange tool


Evaluation Results

Method	AUC	CA	F1	Precision	Recall
Naive Bayes	0.824	0.596	0.585	0.596	0.596
Random Forest	0.999	0.982	0.982	0.982	0.982

Figure 4.1.11 Result

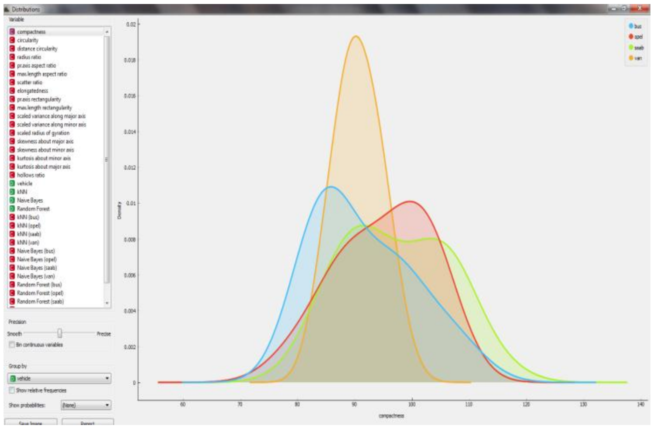


Figure 4.1.12 over all chart

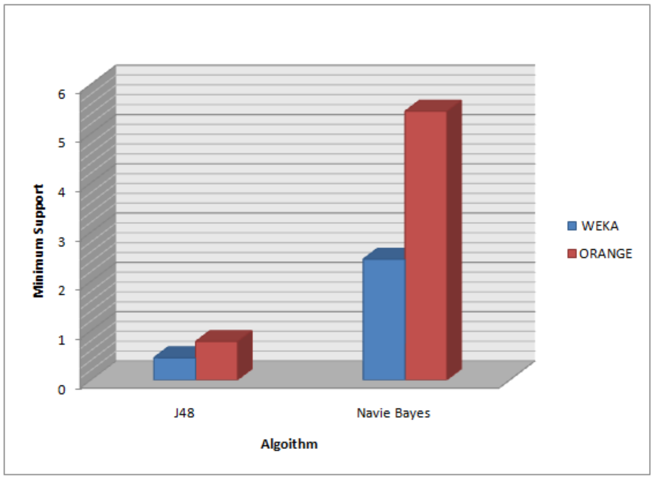
4.2 MINIMUM SUPPPORT

Table 4.2.1 showing minimum support for mushroom dataset

TOOL	J48 Minimum Support	Navie Bayes Minimum Support

WEKA	0.45	2.45
ORANGE	0.78	5.45

Chart 4.2.1 showing minimum support for mushroom dataset

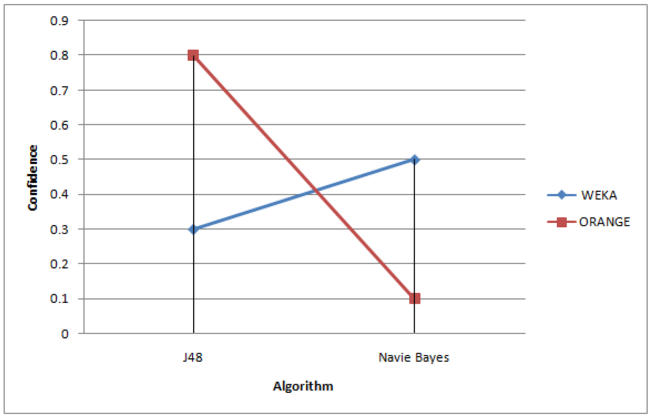


1) *4.2.2 CONFIDENCE*

TABLE 4.2.2 Showing Confidence For Mushroom Dataset

TOOL	J48 CONFIDENCE	Navie Bayes CONFIDENCE
WEKA	0.3	0.5
ORANGE	0.8	0.10

CHART 4.2.2 showing confidence for mushroom dataset



4.2.3 EXECUTION TIME SECOND

Experiments are performed on the mushroom datasets. Machine with configuration of windows 7 system and 2-GB of RAM is used. The results were compared to experiments with WEKA & Orange implementations of J48 and Navies Bayes the techniques run to ensure that the results.

Table 4.2.3 Execution Time Second

TOOL	J48 Execution Time Second	Navie Bayes Execution Time Second
WEKA	0.03	0.15
ORANGE	0.09	0.12

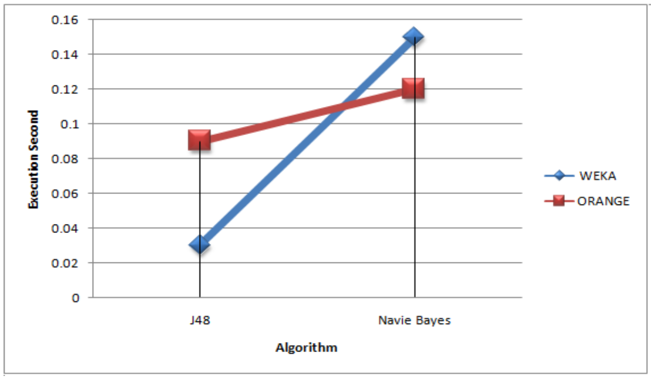


Figure 4.1.13: Representing Dataset measured in Execution Second (ms)

V. CONCLUSION

Discovery of association rules is the most successful and important task in data mining and its goal is to discover all the frequent patterns in the data set. The algorithm does not need to repeatedly scan the database when discovering frequent item sets, therefore can greatly improve the efficiency of data mining.The experiments have been performed using the weka and orange tool. Mushroom data set have been taken from UCI repository having 8124 instances and 22 attributes. J48 is a simple classifier technique to make a decision tree, efficient result has been taken from dataset using weka tool in the experiment. The experiments results shown in the study are about classification minimum support, confidence level and execution time, analysis.J48 gives more classification accuracy for all class in dataset having good result. The result in the study on these mushroom datasets also shows that the efficiency and accuracy of j48 is good.

REFERENCES

[1] Huaifeng Zhang, Yanchang Zhao, Longbing Cao and Chengqi Zhang, “Combined Association Rule Mining”, PAKDD 2008, LNAI 5012, pp. 1069-1074, 2008 © Springer- Verlag Berlin Heidelberg 2008

[2] Pratima Gautam and K.R. Pardasani, “Algorithm for Efficient Multilevel Association Rule Mining” In (IJCSE) International Journal on Computer Science and Engineering, Volume 02, No. 05, 1700-1704, 2010.

[3] Xunwei Zhou and Hong Bao ,” Mning Double-Connective Association Rules from Multiple Tables of Relational Databases “ In IEEE,2008

[4] Raja Tlili and Yahya Slimani, “Executing Association Rule Mining Algorithm under a Gird Computing Environment” In PADTAD,July 2011.

[5] Anis Suhailis Abdul Kadir, Azuraliza Abu Bakar and Abdul Razak Hamdan, “Frequent Absence and Presence Itemset for Negative Association Rule Mining “, IEEE,2011.

[6] Guimei Liu, Haojun Zhang and Limsoon Wong, “Controlling False Positives Iin Association Rule Mining” In Proceedings of the VLDB Endowment ACM,2011.

[7] Somboon Anekritmongkol and M. L. Kulthon Kasamsan , “ The Comparative of Boolean Algebra Compress and Apriori Rule Techniques for New Theoretic Association Rule Mining Model” In IEEE,2009

[8] Salama, G.I.; Abdelhalim, M.B.; Zeid, M.A., "Experimental comparison of classifiers for breast cancer diagnosis," Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on , vol., no., pp.180,185, 27-29 Nov.,2012.

[9] S. Moertini Veronica ,”Towards The Use Of C4.5 Algorithm For Classifying Banking Dataset”,Integeral Vol 8 No 2,October 2013.

[10] UM, Ashwinkumar, and Anandakumar KR. "Predicting Early Detection of Cardiac and Diabetes Symptoms using Data Mining Techniques.",IEEE,pp:161-165,2011

[11] Geetha Ramani R, Lakshmi Balasubramanian, and Shomona Gracia Jacob. "Automatic prediction of Diabetic Retinopathy and Glaucoma through retinal image analysis and data mining techniques." In Machine Vision and Image Processing (MVIP), 2012 International Conference on, pp. 149-152. IEEE, 2012

[12] Sugimoto, Masahiro, Masahiro Takada and Masakazu Toi. "Comparison of robustness against missing values of alternative decision tree and multiple logistic regression for predicting clinical data in primary breast cancer." In Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE, pp. 3054-3057. IEEE, 2013.

[13] Hussein Asmaa S,Wail M. Omar, Xue Li, and Modafar Ati. "Efficient Chronic Disease Diagnosis prediction and recommendation system." In Biomedical Engineering and Sciences (IECBES), 2012 IEEE EMBS Conference on, pp. 209-214. IEEE, 2012.

[14] Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[15] Korting, Thales Sehn. "C4. 5 algorithm and Multivariate Decision Trees." Image Processing Division, National Institute for Space Research--INPE.

[16] Guo, Yang, Guohua Bai, and Yan Hu. "Using Bayes Network for Prediction of Type-2 Diabetes." In Internet Technology And Secured Transactions, 2012 International Conferece For, pp. 471-472. IEEE, 2012.