JCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol.6, Special Issue.11, Dec 2018 E-ISSN: 2347-2693

Association Rule Mining Classification using J48 & Navie Bayes

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— Classification is an important data mining technique based on machine learning with broad applications. It classifies various kinds of data and used in nearly every field of our life. Classification is used to classify every item in a set of data into one of predefined set of classes or groups. This paper describes the performance analysis of Naïve Bayes and J48 classification algorithm based on the correct and incorrect instances of data classification. Naive Bayes algorithm is based on probability and j48 algorithm is based on decision tree. In this paper we compare and 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. The experiments results shown in this paper are about 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 tools also show that the efficiency and accuracy of J48 than Naive Bayes is good.

Keywords: *Data mining, Weka 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.

Association Rule Mining he notion of mining association rules are as follows . In the data mining, the Association rule mining is introduced in to identify unknown facts in huge datasets and drawing inferences on how a subset of items influences the presence of another subset. Let S= {S1, S2, S3............ Sn} be a universe of Items and T= {T1, T2, T3..........Tn} is a set of transactions. Then expression X => Y is an association rule where X and Y are itemsets and X ∩ Y=Φ. Here X and Y are called antecedent and consequent of the rule respectively. This rule holds support and confidence, support is a set of transactions in set T that contain both X and Y and confidence is percentage of transactions in T containing X that also contain Y. An association rule is strong if it satisfies user-set minimum support (minsup) and minimum confidence (minconf) such as support ≥ minsup and confidence ≥ minconf. An association rule is frequent if its support is such that support ≥ minsup

II LITERATURE SURVEY

Huaifeng 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.

Hong 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 highest

power 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 OVERVIEW OF DATAMINING

Data Mining is the innovation of hidden information found in large quantities of data and can be viewed as a step in the knowledge discovery process (Fayyad 1996).Data mining defined as a set of computer-assisted techniques designed to automatically mine big volumes of integrated data for new, hidden or unexpected information, or interesting patterns. With small set of data, traditional statistical analysis can be efficiently used.

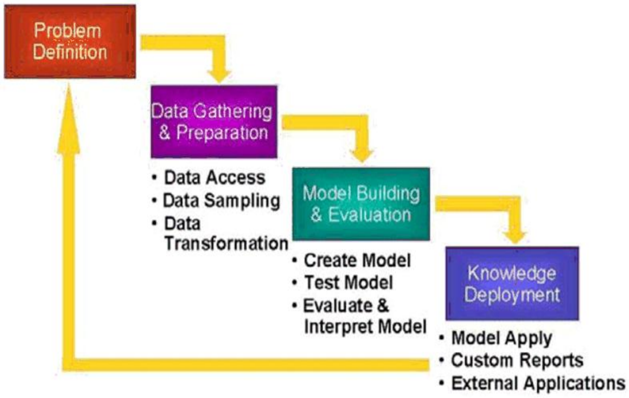


Figure 3.1 An overview of steps that compose KDD process

3.2 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

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

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.

3.3 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. Model generated by decision tree helps to predict new instances of data.

```
Algorithm [1] J48:
INPUT
D // Training data
OUTPUT
T // Decision tree
DTBUILD (*D)
{
T = Null;
T = Create root node and label with splitting attribute;
T = Add oval to root node for every split predicate and tag;
For each arc do
D = Database created by applying splitting g predicate to D;
If stop point reached for this path, then
T'= Create leaf node and label with appropriate class;
Else
T' = DTBUILD (D);
T = Add T' to arc;
```

While building tree J48 ignores the missing value. J48 allows classification via either decision n tree or rules generated from them.

IV. EXPERIMENT AND RESULTS

The proposed system has been implemented using weka tool environment. It can be executed on windows. The results are obtained as follows after execution Mushroom data set is presented.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 Navie Bayes. It takes less time to generate frequent item sets.

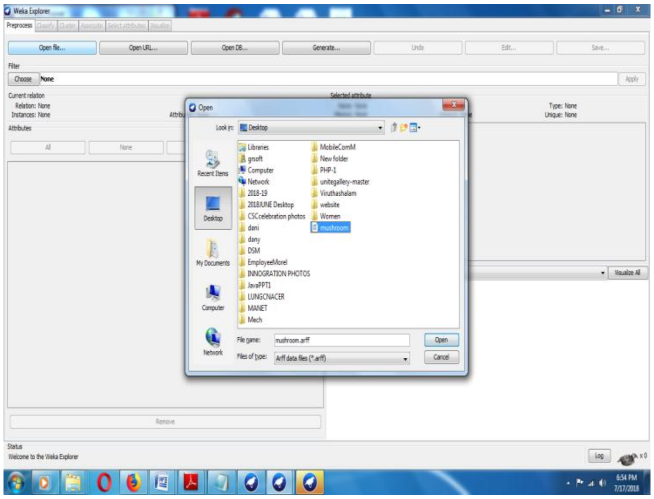
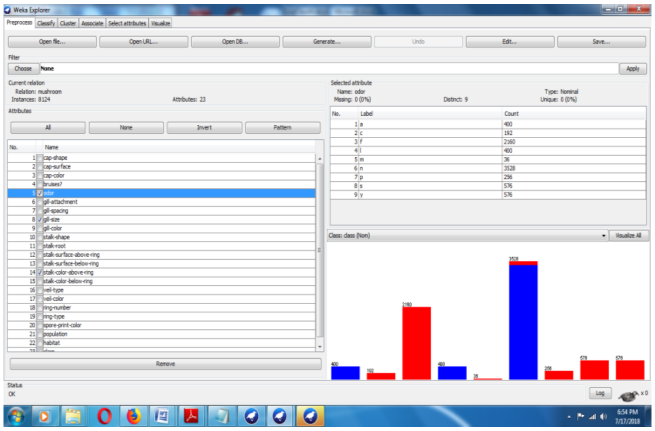


Figure 4.1.1 Upload Mushroom Dataset



Current relation
Relation: mushroom
Instances: 8124
Attributes: 23

Selected attribute
Name: odor
Missing: 0 (0%)
Distinct: 9
Type: Nominal
Unique: 0 (0%)

No.	Name
1	cap-shape
2	cap-surface
3	cap-color
4	bruises?
5	odor
6	gill-attachment
7	gill-spacing
8	gill-size
9	gill-color
10	stalk-shape
11	stalk-root
12	stalk-surface-above-ring
13	stalk-surface-below-ring
14	stalk-color-above-ring
15	stalk-color-below-ring
16	veil-type
17	veil-color
18	ring-number
19	ring-type
20	spore-print-color
21	population
22	habitat

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

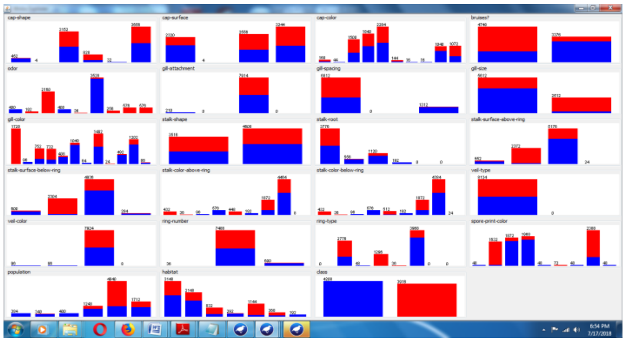
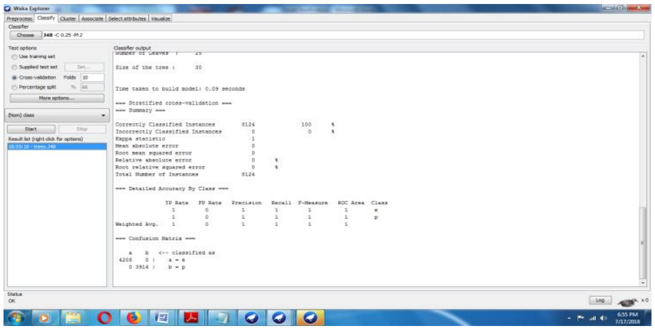
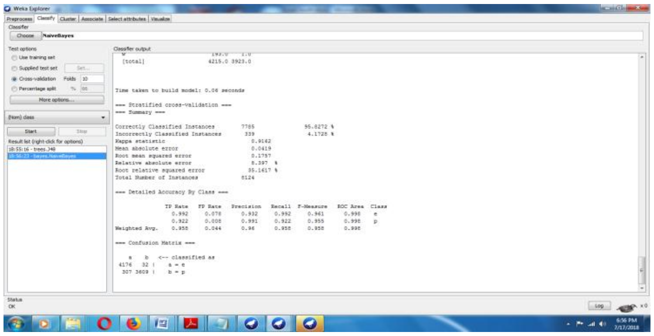


Figure 4.1.3 Over All Chart



```
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



```
  w                      193.0    1.0
  [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                          0.9142
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

4.2 MINIMUM SUPPPORT

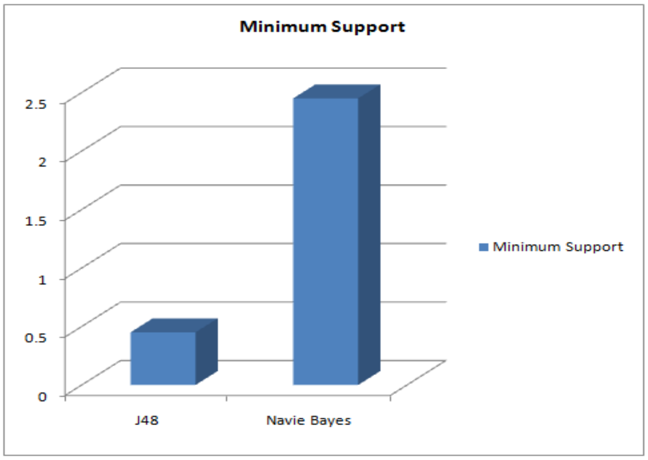
Support is an indication of how frequently the itemset appears in the dataset. The support of X with respect to T is defined as the proportion of transactions t in the dataset which contains the itemset X.

$$\mathrm{supp}(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$$

Table 4.2.1 showing minimum support for mushroom dataset

Algorithm	Minimum Support
J48	0.45
Navie Bayes	2.45

Chart 4.2.1 showing minimum support for mushroom dataset



1) 4.2.2 CONFIDENCE

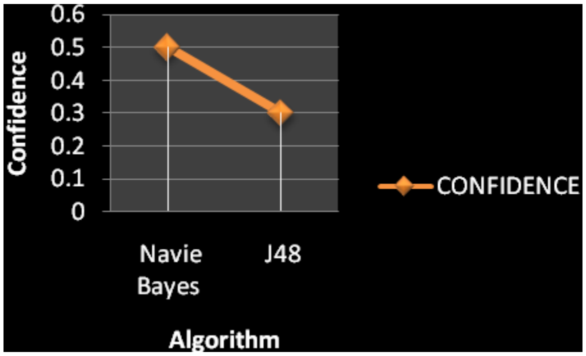
Confidence is an indication of how often the rule has been found to be true. The *confidence* value of a rule, X➔Y, with respect to a set of transactions T, is the proportion of the transactions that contains X which also contains Y. Confidence is defined as:

$$\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X)$$

TABLE 4.2.2 Showing Confidence For Mushroom Dataset

ALGORITHM	CONFIDENCE
Navie Bayes	0.5
J48	0.3

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 implementations of J48 and Navies Bayes the techniques run to ensure that the results.

Table 4.2.3 Execution Time Second

ALGORITHMS	TIME TAKEN (IN MIL SECS)
Navie Bayes	0.15
J48	0.03

Figure : Representing Dataset measured in Execution Second(ms)

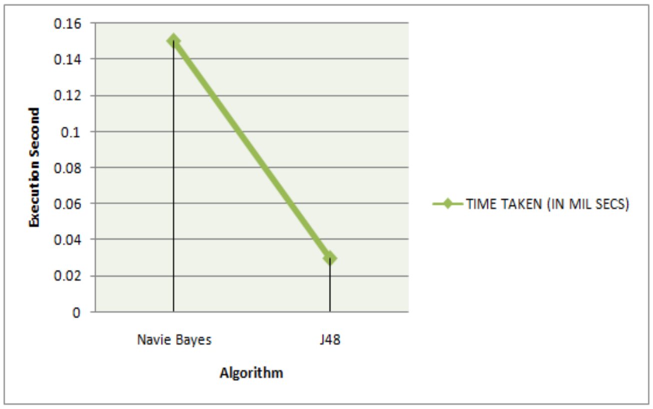


Fig.4.1.6

V. CONCLUSION

In this paper the algorithms that are dealing the association rule mining with Classification algorithms are compared and analyzed. The proposed method uses mushroom dataset. The experiments have been performed using the weka tool. Mushroom data set have been taken from UCI repository having 8124 instances and 22 attributes. J48 is a easy classifier algorithm 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 confidence level and execution time, minimum support analysis J48 gives more classification accuracy for all class in dataset having good result. The result in the study on these datasets also shows that the efficiency and accuracy of j48 is good. Classification technique of data mining is useful in every domain. Here we find out the best as Weka tool its execution time will be faster as compare to Weka tool.

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.