An Empirical Study of Naive Bayes Classification, K-Means Clustering and Apriori Association Rule for Supermarket Dataset

Aishwarya. C. M
Department of computer science and Engineering
A.I.T College, Chikmagalur

Arpitha. D. A
Department of computer science and Engineering
A.I.T College, Chikmagalur

Prerana. B. R
Department of computer science and Engineering
A.I.T College, Chikmagalur

Rachana. C. L
Department of computer science and Engineering
A.I.T College, Chikmagalur

Karthikeyan. S. M
Asst. Prof. Department of computer science and Engineering
A.I.T College, Chikmagalur

Abstract- **Data mining is the process of obtaining knowledge from the large quantity of data. Generally, it is the process of interpreting data from different perspectives and summarizing it into useful information. It analyses the most useful and familiar data mining tasks like classification, clustering and association rules used by the Machine learning system, mainly in the artificial intelligent systems. The paper is mainly concerned with practical use of classification, clustering and association rule that is applied to the dataset of supermarket. In this we are studying the result of the three data mining tasks. The paper predicts the performance evaluation based on the incorrect and correct instances of data using the Naive Bayes, K-means and Apriori algorithm. The analysis of classification algorithm, clustering algorithm and association rule is done by WEKA tool.**

Keywords- Dataset, Weka Analysis, Classification Algorithms, Clustering Algorithms, Association Rules

I. INTRODUCTION

Data mining is developing vastly in various fields. Data mining is adopted for several use and designed for different database [8]. The data mining task includes classification, clustering, association rule extraction, regression and visualization [7]. Classification is a data mining task that assigns items in a collection to target categories, classes. The goal of classification is to accurately obtain the target class for each case in the data [8]. Clustering is a process of partitioning a data set into set of meaningful sub-classes called clusters. Help users understand the natural grouping or structure in a data set. Association rule is procedure which is meant to find frequent patterns, correlations, associations, or casual structures from data sets found in various kinds of databases such as relational databases, transactional databases and other forms of data repositories. For the execution of classification, clustering and association rule we have used WEKA tool. WEKA is a general collection of machine learning software written in Java, developed at the University of Waikato in New Zealand. Weka is a workbench that contains a collection of visualization tools and algorithm for data analysis and predictive modelling.

Throughout the discussion we try to understand some of the tests, analysis of classification, analysis and association rule. We make use of Weka tool to implement the different data mining tasks for the dataset supermarket. Here, we discuss about the content of data and fields which are related to the data set. The discussion is followed as first we discuss the nature of classification algorithm Naive Bayes, clustering algorithm K-means and association rule using Apriori algorithm . The next discussion is all about analysis of classification, clustering and association rule based on supermarket dataset. To analyse the supermarket datasets we use algorithms, which include Naive Bayes [4], K-means and Apriori algorithm. Finally the research results in the study of supermarket data set based on the algorithms used in the Weka tool.

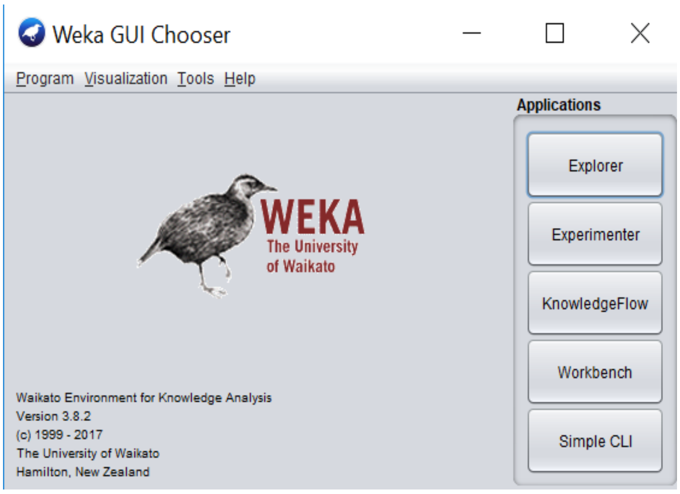
II. METHODOLOGY

We have used Weka tool for the analysis of three different data mining tasks. The algorithms of the data mining task can be directly applied to a dataset. It is also well suited for developing new machine learning schemes. The data that is used for Weka should be made into the arff format and the field should have the extension dot arff (.arff). The arff works with three sections they can be categorised into Relational section, Attribute section and the data. Whereas the data types of arff are classified into Nominal and Numeric.

A. *WEKA GUI Chooser*

The WEKA GUI Chooser is the starting point for running the applications. There will be a choice between the command line interface (CLI), the Experiment, the Explorer and Knowledge flow. In this context we make use of explorer that gives access to all features of Weka using menu selection and form filling.

The figure1 shows the WEKA GUI chooser for commencing the main application.

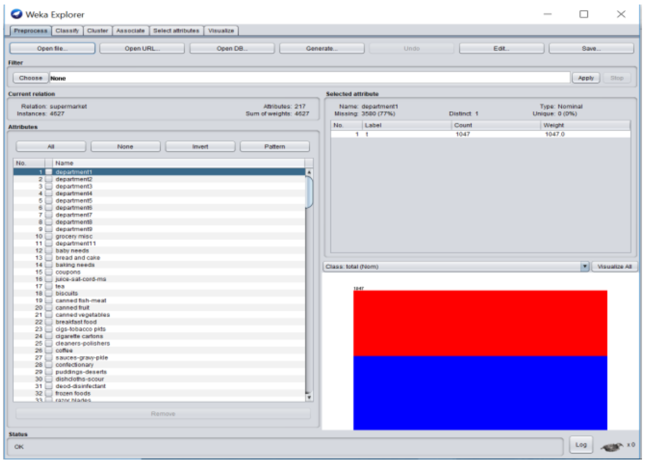


Waikato Environment for Knowledge Analysis
Version 3.8.2
(c) 1999 - 2017
The University of Waikato
Hamilton, New Zealand

Fig1: WEKA GUI chooser

B. *Supermarket dataset*

The supermarket dataset is a dataset of point of sale information. The data is nominal and each instance represents a customer transaction at a supermarket, the product purchased and the departments involved. The performance of a comprehensive set of classification algorithms has been analysed. The dataset contains 4,627 instances and 217 attributes. The data is denormalized. Each attribute is binary and either has a value or no value. There is nominal class attribute called "total" that indicates whether transaction was less than $100 (low) or greater than $100 (high). The figure2 shows the supermarket data set opened in WEKA tool.



Filter: None

Relation: supermarket
Instances: 4627
Attributes: 217
Sum of weights: 4627

Name: department1
Missing: 3580 (77%)
Distinct: 1
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	1	1047	1047.0

No.	Name
1	department1
2	department2
3	department3
4	department4
5	department5
6	department6
7	department7
8	department8
9	department9
10	grocery misc
11	department11
12	baby needs
13	bread and cake
14	baking needs
15	coupons
16	juice-sat-cord-ms
17	tea
18	biscuits
19	canned fish-meat
20	canned fruit
21	canned vegetables
22	breakfast food
23	cigs-tobacco pkts
24	cigarette cartons
25	cleaners-polishers
26	coffee
27	sauces-gravy-pkle
28	confectionary
29	puddings-deserts
30	dishcloths-scour
31	deod-disinfectant
32	frozen foods
33	razor blades

OK

Fig2: Supermarket datasets open in WEKA

III. DATA MINING TASK

A. *CLASSIFICATION ALGORITHM*

There are many classification algorithms available on WEKA tool but in the paper we have selected only one classification algorithm. Analysis of classification algorithm is the assembling of data in given classes.

- *Naive Bayes*

The Naive Bayes algorithm is a simple probabilistic classifier that determines a set of possibilities by counting the constancy and combination of values in given data set [3]. This algorithm is also used in machine learning systems to conclude the new data or testing data, and it is based on the "Bayes" theory [4]. The application of this algorithm is performed by Weka tool, which provide opportunity to implement the above mentioned algorithm by using the estimator, for the numeric attributes.

B. CLUSTERING ALGORITHM

Clustering is a technique in which a given data set is divided into groups called clusters in such a manner that the data points that are similar lie together in one cluster. Clustering plays an important role in the field of data mining due to the large amount of data set. We have used K-means clustering to obtained clustered instances.

- *K-means algorithm*

K-MEANS Clustering intends to partition n objects into k clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly k different clusters of greatest possible distinction. The best number of clusters k leading to the greatest separation is not known as priori and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function.

C. ASSOCIATION RULES

This is a procedure which is meant to find the request pattern, correlations, associations or casual structures from data sets. Given a set of transactions, association rule mining aims to find the rules which enable us to predict the occurrence of a specific item based on the occurrences of the other items in the transaction.

- *Apriori algorithm*

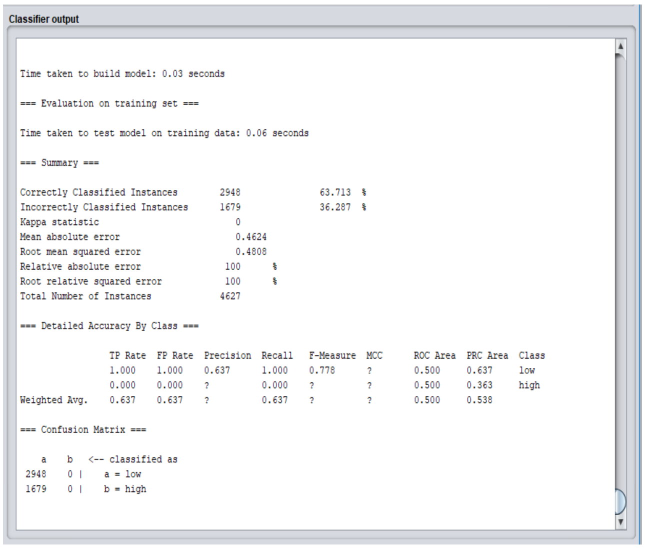
This is the most well known association rule learning method because it may have been first and it is very efficient. In principle the algorithm is quite simple. It builds up attribute-value (item) sets that maximize the number of instances that can be explained (coverage of the dataset). The search through item space is very much similar to the problem faced with attribute selection and subset search.

IV. SUPERMARKET DATASET ANALYSIS

- *Analysis using Naive Bayes Algorithm*

The initial analysis of this dataset is performed by default parameters that are provided by weka tool. The classification is performed by choosing "use training set".

Which specifies that out of 4627 instances the correctly classified instances are 2948 (63.713%) and incorrectly classified instances are 1679 (36.287%). The Fig3 represents the Naive Bayes Algorithm outcome.

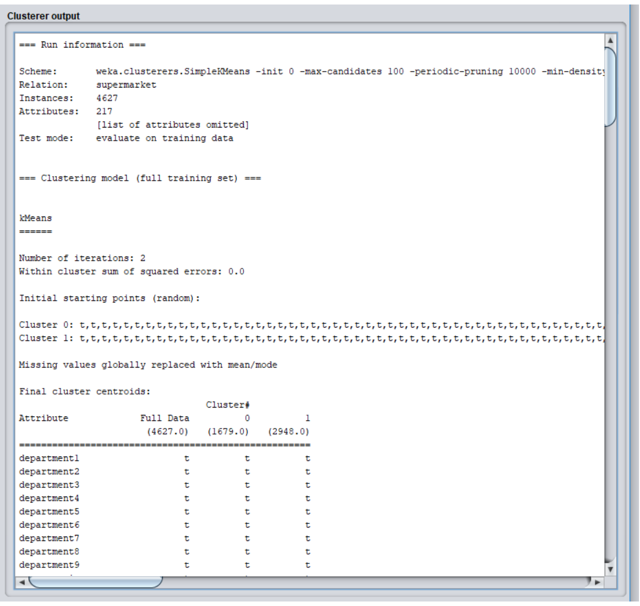


```
Time taken to build model: 0.03 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.06 seconds

=== Summary ===

Correctly Classified Instances        2948               63.713 %
Incorrectly Classified Instances      1679               36.287 %
Kappa statistic                          0
Mean absolute error                      0.4624
Root mean squared error                  0.4808
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             4627

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.637      1.000    0.778      ?        0.500     0.637     low
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.363     high
Weighted Avg.    0.637    0.637    ?          0.637    ?          ?        0.500     0.538

=== Confusion Matrix ===

    a    b   <-- classified as
 2948    0 |    a = low
 1679    0 |    b = high
```

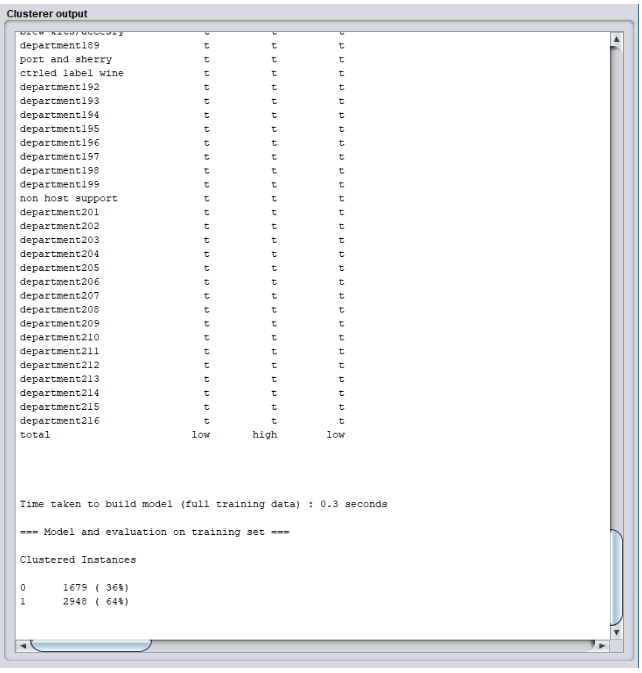
Fig3: Output of Naïve Bayes Algorithm

B. *Analysis using K-means algorithm*

The K-means algorithm is a clustering algorithm, here training data is used for clustering it generates two clustered instances of 1679 (36%) and 2948 (64%). The total number of iterations done are two.



```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-densit
Relation:     supermarket
Instances:    4627
Attributes:   217
              [list of attributes omitted]
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 0.0

Initial starting points (random):

Cluster 0: t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,
Cluster 1: t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,t,

Missing values globally replaced with mean/mode

Final cluster centroids:
                                  Cluster#
Attribute             Full Data          0          1
                       (4627.0)   (1679.0)   (2948.0)
=====================================================
department1                   t          t          t
department2                   t          t          t
department3                   t          t          t
department4                   t          t          t
department5                   t          t          t
department6                   t          t          t
department7                   t          t          t
department8                   t          t          t
department9                   t          t          t
```



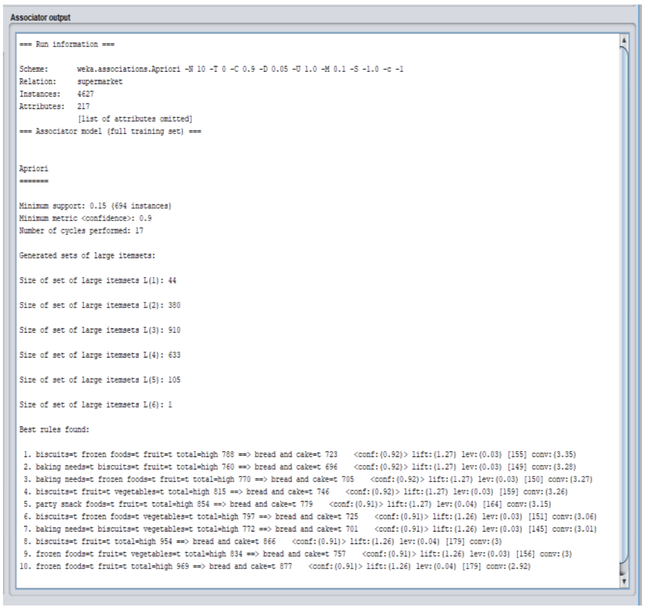
```
brew kits/accesry            t          t          t
department189                t          t          t
port and sherry              t          t          t
ctrled label wine            t          t          t
department192                t          t          t
department193                t          t          t
department194                t          t          t
department195                t          t          t
department196                t          t          t
department197                t          t          t
department198                t          t          t
department199                t          t          t
non host support             t          t          t
department201                t          t          t
department202                t          t          t
department203                t          t          t
department204                t          t          t
department205                t          t          t
department206                t          t          t
department207                t          t          t
department208                t          t          t
department209                t          t          t
department210                t          t          t
department211                t          t          t
department212                t          t          t
department213                t          t          t
department214                t          t          t
department215                t          t          t
department216                t          t          t
total                      low       high        low




Time taken to build model (full training data) : 0.3 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      1679 ( 36%)
1      2948 ( 64%)
```

Fig4: Output of K-means Algorithm

C. *Analysis using Apriori Algorithm*

The algorithm presented 10 rules learned from the supermarket dataset. The algorithm is configured to stop at 10 rules by default, for more rules we can configure it. Association output has no rule with coverage less than 0.91. Here are few observations:

- We can see that all presented rules have a consequent of “bread and cake”.
- All presented rules indicate a high total transaction amount.
- “biscuits” an “frozen foods” appear in many of the presented rules.



```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     supermarket
Instances:    4627
Attributes:   217
              [list of attributes omitted]
=== Associator model (full training set) ===



Apriori
=======

Minimum support: 0.15 (694 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 44

Size of set of large itemsets L(2): 380

Size of set of large itemsets L(3): 910

Size of set of large itemsets L(4): 633

Size of set of large itemsets L(5): 105

Size of set of large itemsets L(6): 1

Best rules found:

 1. biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723    <conf:(0.92)> lift:(1.27) lev:(0.03) [155] conv:(3.35)
 2. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696    <conf:(0.92)> lift:(1.27) lev:(0.03) [149] conv:(3.28)
 3. baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t 705    <conf:(0.92)> lift:(1.27) lev:(0.03) [150] conv:(3.27)
 4. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746    <conf:(0.92)> lift:(1.27) lev:(0.03) [159] conv:(3.26)
 5. party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779    <conf:(0.91)> lift:(1.27) lev:(0.04) [164] conv:(3.15)
 6. biscuits=t frozen foods=t vegetables=t total=high 797 ==> bread and cake=t 725    <conf:(0.91)> lift:(1.26) lev:(0.03) [151] conv:(3.06)
 7. baking needs=t biscuits=t vegetables=t total=high 772 ==> bread and cake=t 701    <conf:(0.91)> lift:(1.26) lev:(0.03) [145] conv:(3.01)
 8. biscuits=t fruit=t total=high 954 ==> bread and cake=t 866    <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(3)
 9. frozen foods=t fruit=t vegetables=t total=high 834 ==> bread and cake=t 757    <conf:(0.91)> lift:(1.26) lev:(0.03) [156] conv:(3)
10. frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877    <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(2.92)
```

Fig5: Output of Apriori Algorithm

V. CONCLUSION

The purpose of a data mining tasks is normally either to create a descriptive model or a predictive model. This approach analyses the application of three different algorithms of data mining task in the data set. Naive Bayes algorithm is intended primarily for the work with nominal attributes. The performance depends on the classification algorithm that is adopted. The K-means algorithm results in 2 cluster instances and the Apriori algorithm gives us the information that if itemset is frequent then all its subset is frequent and the generates several rules. It is more efficient to use an algorithm like Apriori rather than deducing rules by hand.

ACKNOWLEDGEMENT

We would like to thank Mr. Karthikeyan S M, Asst. Professor, Department of Computer Science and Technology, Adichunchanagiri Institute of Technology Chickamagalur for assistance we have received in writing this paper.

REFERENCES

[1] Thangaraju, R. Mehla, Analysis of KStar Classifier over Liver Disease, International Journal of Advanced Research in Computer science Engineering & Technology (IJARCET) Volume 4 Issue 7,July 2015

[2] Ho Tin Kham. “Random subspace Methods for Constructing Decision Forest”

[3] George Dimytoglou, James Adam, and Carol M. Jhim, “Comparisons of C4.5 and Naive Bayes Classification for the Analysis of Lung Cancer Survivabilities”

[4] Olivier C. Fhran, kois and Philip Leray Study of the Tree Augmented Naive Bayes Classification from deficient datasets LITIS. Sain-Etienne-De-Rouary, France.

[5] Jangtao Ron*,Sau Dan Le, Xianlo Chn , Ben Ka, Renold Chenk and David Cheunk Naive Bayer Classification of incalculable Data Department of Computer Engineering, Son Yaat-son University, Guangzhou, China

[6] Lio Bhreman, Jerom Fridman, Richerd Olshan, Charle Stone “Regression Tree” (Wardsworth).

[7] Ghopi Gandi, Rohith Shreevastav Modified k-methods algorithms for analysis and application to increase scalabilities and efficiencies for larger datasets.