Mining Association Rules using R Environment

Deepa U. Mishra
Asst. Professor (Comp. Engg.)
NBNSSOE , Pune

Nilam K. Kadale
Asst. Professor (Comp. Engg.)
NBNSSOE. Pune

ABSTRACT

R is an unified collection of software attachments for performing various operations on data and graphical display. R has become a preferred platform for statistical analysis.

The R add-on package **arules** implements the basic infrastructure for creating and manipulating transaction databases and basic algorithms to efficiently find and analyze association rules. Compared to other tools, the **arules** framework is fully integrated, implements the latest approaches and has the vast functionality of R for further analysis of found patterns at its disposal. The "apriori" function, provided by the arules package, is used for mining association. Apriori Algorithm is the most popular algorithm for mining association rules. Association rule of data mining is employed in all tangible applications of business and industry. Objective of taking Apriori is to find frequent item sets and to disclose the unreleased information. This paper encompasses the use of association rule mining in extracting patterns that occur frequently within a dataset and showcases the importance of the Apriori algorithm in mining association rules from a dataset containing student information.

Keywords

Data Mining, The R Project, Association Rules, Apriori Algorithm.

1. INTRODUCTION

Data Mining is an elaborated process of analyzing enormous amounts of data and picking out the meaningful information. It refers to extracting or mining knowledge from enormous amounts of data[1,2]. The data sources incorporate databases, data warehouses, the World Wide Web(WWW), other information repositories, or data that are included in the system dynamically [3,4]. Association Rule in Data Mining plays a vital role in the process of mining data for frequent itemsets. Mining frequent itemsets is a popular method for finding associated items in databases. Association rules is one of the most successful data mining techniques. Frequent patterns occur periodically in the data. Patterns can contain itemsets, sequences and subsequences. A frequent itemset associates to a set of items that usually appear together in a transactional data set[5]. It involves the following steps: cleaning and integrating data from data sources like databases, flat files, pre-processing of selection and transformation of target data, mining the necessary knowledge and ultimately assessment and presentation of knowledge. A data mining algorithm is said to be absolute if it mines all interesting patterns. User-procured constraints and interestingness measures are to be used to focus the search for interesting patterns rather than searching for all possible patterns.

The R Project allows for a multitudinous of specialized packages to be installed and employed by its users as required. These contain packages and functions for model building and predictive analytics. It has an efficient data manipulation and storage capability. It includes a collection of operators for performing operations on arrays, particularly matrices and an extensive collection of intermediate tools for analysis of data. It provides graphical capabilities for analysis and display of data.

2. RELATED DEFINITION

Association Rule: Association rule of data mining includes picking out the unknown inter-dependence of the data and interpreting the rules between those items [6]. A rule is defined as an implication of the form A=>B, where A∩ B≠¢. The left-hand side of the rule is known as antecedent and the right-hand side of the rule is known as consequent.

Support: I = { i1,i2,i3, … , im} is a collection of items. Let T be a collection of transactions associated with the items. Each transaction is provided an identifier TID [7]. Association rule A=>B is such that A ϵ I, B ϵ I. A is termed as Premise and B is termed as Conclusion. The Support (S), is defined as the ratio of transactions in the data set that contains the itemset. Support(X=>Y) = Support (XUY) = P(XUY).

Confidence: The confidence is defined as a conditional probability Confidence (X=>Y) = Support (XUY) / Support(X) = P(Y/X)[8,9].

Lift: is defined as the ratio of the probability that L and R eventuate together to the multiple of the two individual probabilities for L and R.

Lift = Pr(L,R) / Pr(L).Pr(R).

Conviction: is identical to lift, despite it measures the effect of the right-hand-side not being true. It also inverts the ratio. A conviction is calculated as:

conviction = Pr(L).Pr(not R) / Pr(L,R)

3. THE APRIORI ALGORITHM

A realization of frequent pattern matching on the basis of support and confidence dimensions produced excellent results in various fields[10,11]. The apriori pseudocode is given below that uses two parameters , that is a database containing items and a minimum support count value for finding the frequent items

Apriori Algorithm Pseudocode

```
Routine Apriori (T, MinSupport) { where T is a database and MinSupport is the minimum support count

L1= {frequent items};

for (k= 2; Lk-1 != Ǿ; k++) {

Ck= candidates generated from Lk-1

for each transaction t in database do

{

#increment the candidates count in Ck that are contained in t

Lk = candidates in Ck with  MinSupport

}
}
return UKLK ;
}
```

In association rule mining, given a set of item-sets (consider an example, given a collection of retail transactions, each containing individual items brought), Apriori algorithm tries to find the subsets that are typical to at least a minimal number (C) of the item-sets. Apriori implements a 'bottom up' technique, in which frequent subsets are elaborated one item at once (this step is called candidate generation), and the collection of candidates are examined with reference to the data. The algorithm halts when no further successful elaborations are discovered.

Apriori employs breadth-first search and a tree structure for counting the candidate item sets conveniently. It produces k size candidate item-sets from item-sets of size k-1. After this it eliminates the candidate item sets that contain an infrequent sub pattern. According to the downward closure lemma, the candidate item set encloses all frequent item sets of length k. After that, it inspects the transaction database for resolving frequent item amidst the candidate item sets.

4. IMPLEMENTATION OF MINING USING arules PACAKGE

This R package provides the infrastructure for representing, manipulating and analyzing transaction data and patterns (frequent itemsets and association rules). It also provides interfaces for C applications of the association mining algorithms such as Apriori and Eclat.

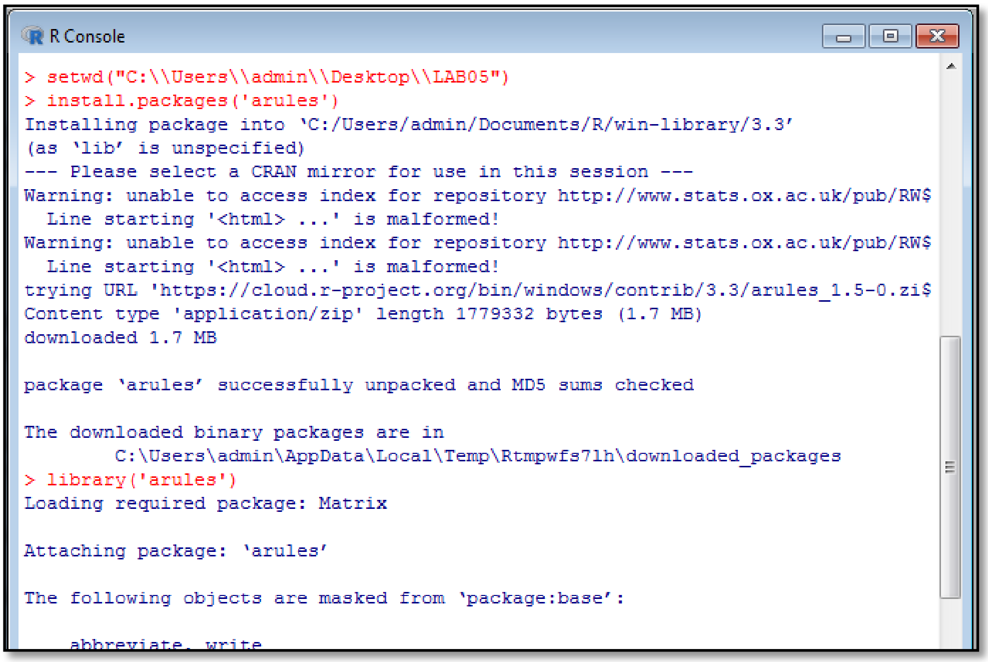
Other packages in the arules family are:

- arulesViz: Visualization of association rules.
- arulesCBA: Classification on the basis of association rules.
- arulesSequences: Mining of frequent sequences. Firstly we have to load the package and the required libraries. The following command is used for installing arules package.

  > install.packages('arules')

  Then a Cran mirror is to be selected_for installing the package. The required package is loaded using the following command as shown in figure 1.

  > library('arules')



```
> setwd("C:\\Users\\admin\\Desktop\\LAB05")
> install.packages('arules')
Installing package into ‘C:/Users/admin/Documents/R/win-library/3.3’
(as ‘lib’ is unspecified)
--- Please select a CRAN mirror for use in this session ---
Warning: unable to access index for repository http://www.stats.ox.ac.uk/pub/RW$
  Line starting '<html> ...' is malformed!
Warning: unable to access index for repository http://www.stats.ox.ac.uk/pub/RW$
  Line starting '<html> ...' is malformed!
trying URL 'https://cloud.r-project.org/bin/windows/contrib/3.3/arules_1.5-0.zi$
Content type 'application/zip' length 1779332 bytes (1.7 MB)
downloaded 1.7 MB

package ‘arules’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\admin\AppData\Local\Temp\Rtmpwfs71h\downloaded_packages
> library('arules')
Loading required package: Matrix

Attaching package: ‘arules’

The following objects are masked from ‘package:base’:

    abbreviate, write
```

Fig 1: Installing arules package

4.1 Read in the Data for Modeling

Transaction List is a special data type function in the “arules” package. The data is read in as a Transaction List using the following statement for the states data, “MBAdata.csv”.txn<-read.transactions ("MBADATA.csv",rm.duplicates=False,format="single",sep=",",cols=c(1,2))

The arguments of the read.transaction functions are listed below:

(i) File : the name of the file .

(ii) format: It is a character string representing the format of the data set.

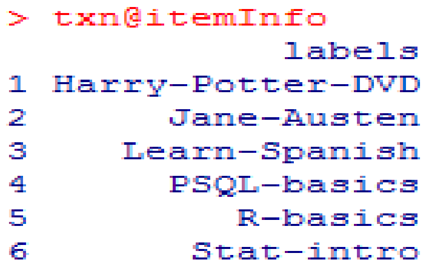
(iii) Sep: It is a character string indicating how fields are distributed in the data file, or NULL (default). For basket format, it can be a regular expression; or else, a single character should be given. The default correlates to white space separators.

(iv) Cols : For the ‘single’ format, cols is a numeric vector of length two giving the numbers of the columns (fields) with the transaction and item ids, respectively. For the ‘basket’ format, cols can be a numeric scalar giving the number of the column (field) with the transaction ids. If cols = NULL

(v) rm.duplicates: It is a logical value indicating if duplicate items must be removed from the transactions.

4.2 Review Transaction data

First the transaction data is inspected as shown in figure 2.



```
> txn@itemInfo
            labels
1 Harry-Potter-DVD
2      Jane-Austen
3    Learn-Spanish
4      PSQL-basics
5         R-basics
6       Stat-intro
```

Fig 2: Transaction data

4.3 Plot Transactions

The "image" function can be used to display a visual representation of the transaction set in which the rows are individual transactions (identified by transaction ids) and the dark squares are items contained in each transaction.

- image(txn)

Figure 3 shows the output of image function. It displays a graph of Transactions vs Items

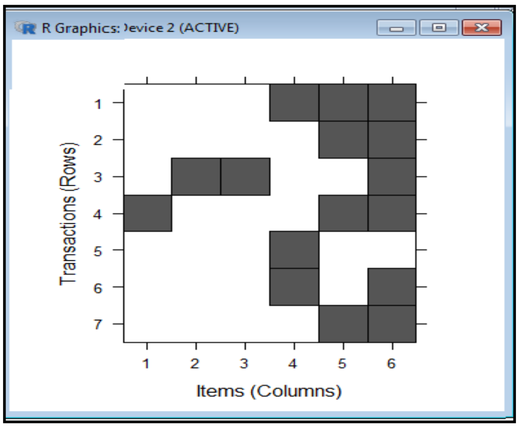


Fig 3: Transactions vs Items

4.4 Mining the Association Rules

The "apriori" function, included in the arules package, is used as follows:

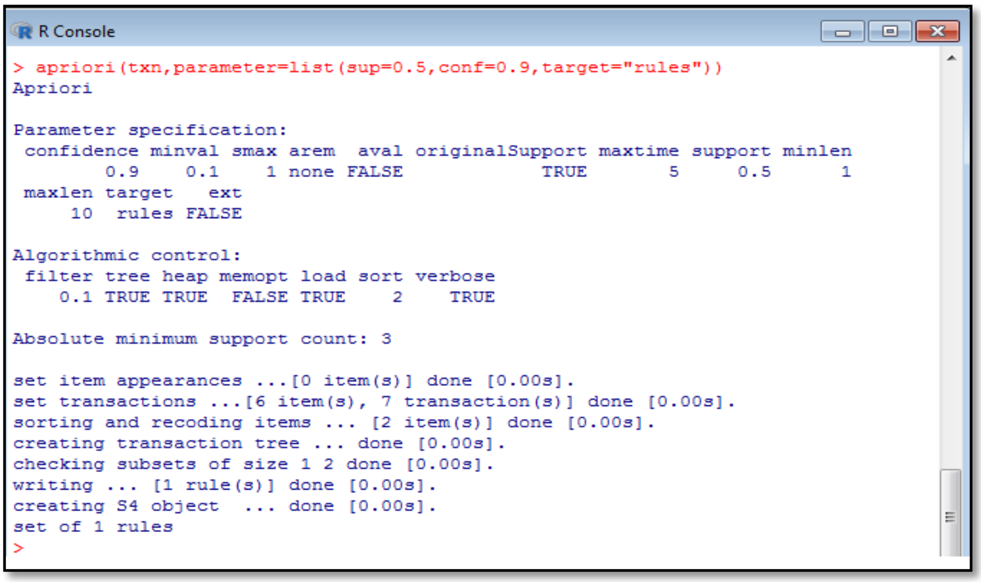
rules <- apriori(File, parameter =list(supp = 0.5, conf =0.9,target ="rules"))

where the arguments are:

data: It is a object of class transactions or any data structure that can be forced into transactions (for example a binary matrix or data.frame).

paramete : It is a named list. The default behavior is to mine rules with support count 0.1, confidence 0.8 and maxlen 5.

The following statement is used for reading the transaction data as shown in figure 4



```
> apriori(txn,parameter=list(sup=0.5,conf=0.9,target="rules"))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
        0.9    0.1    1 none FALSE            TRUE       5     0.5      1
 maxlen target   ext
     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 3

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[6 item(s), 7 transaction(s)] done [0.00s].
sorting and recoding items ... [2 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 done [0.00s].
writing ... [1 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
set of 1 rules
>
```

Fig 4: Reading the Transaction data

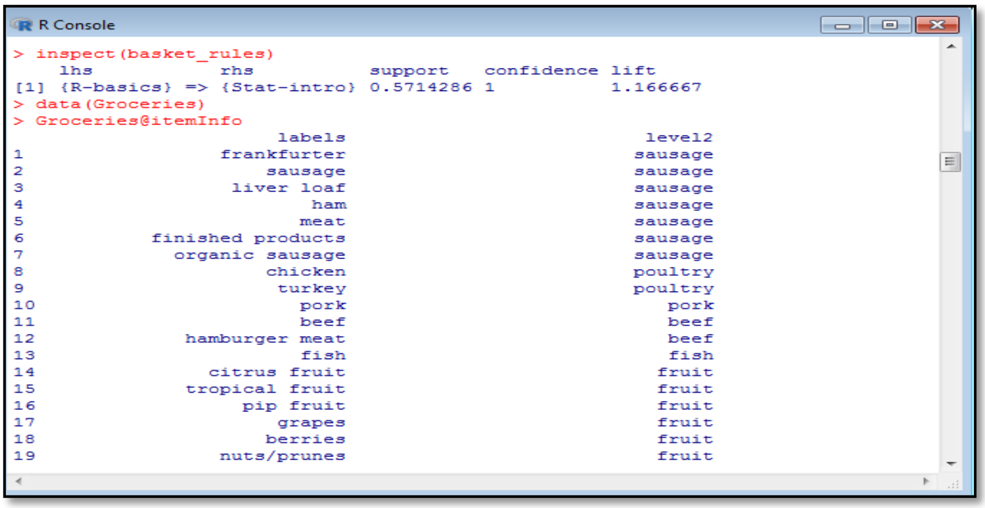
The number of rules generated can be seen in the output and is represented as follows

writing ... [1 rule(s)] done [0.00s]

The rules can be observed using the inspect command.

> inspect(basket_rules)

The output for the inspect command is shown in Figure 5. By reviewing the output it is observed that the generated rule has a support count of 0.571, confidence of the rule is 1 and the lift threshold for the rule is 1.167



```
> inspect(basket_rules)
    lhs           rhs          support   confidence lift
[1] {R-basics} => {Stat-intro} 0.5714286 1          1.166667
> data(Groceries)
> Groceries@itemInfo
                       labels                           level2
1                 frankfurter                          sausage
2                     sausage                          sausage
3                  liver loaf                          sausage
4                         ham                          sausage
5                        meat                          sausage
6           finished products                          sausage
7             organic sausage                          sausage
8                     chicken                          poultry
9                      turkey                          poultry
10                       pork                             pork
11                       beef                             beef
12             hamburger meat                             beef
13                       fish                             fish
14               citrus fruit                            fruit
15             tropical fruit                            fruit
16                  pip fruit                            fruit
17                     grapes                            fruit
18                    berries                            fruit
19                nuts/prunes                            fruit
```

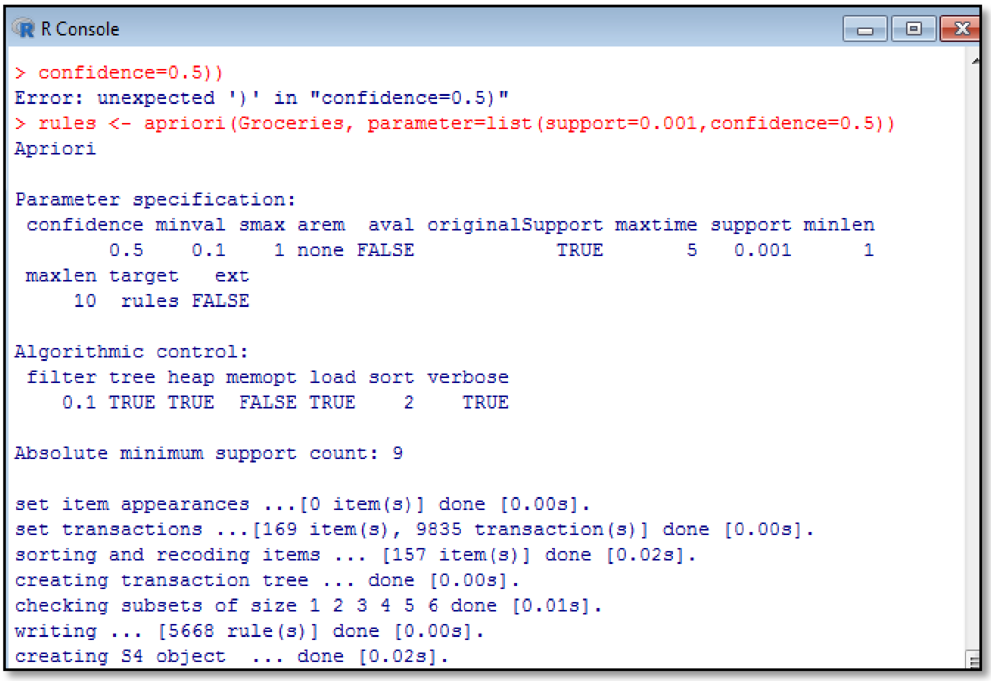
Fig 5. Rule generation

4.5 Reading from the Groceries dataset

Using the standard data set, available with the “arules” package that is “Groceries”. The Groceries data set contains one month of real-world transaction data from a normal local grocery store. The data set includes 9835 transactions and the items are combined to 169 categories. Data can be read in the data set and the item information can be observed.

4.6 Mining the rules for the Groceries Data

The “apriori” function, is applied on the Groceries data set for mining the association rules. The rules are mined by setting a specific value of support and confidence . The value of support is set to .001 and confidence is 0.5. Figure 6 shows mining of association rules using apriori function.

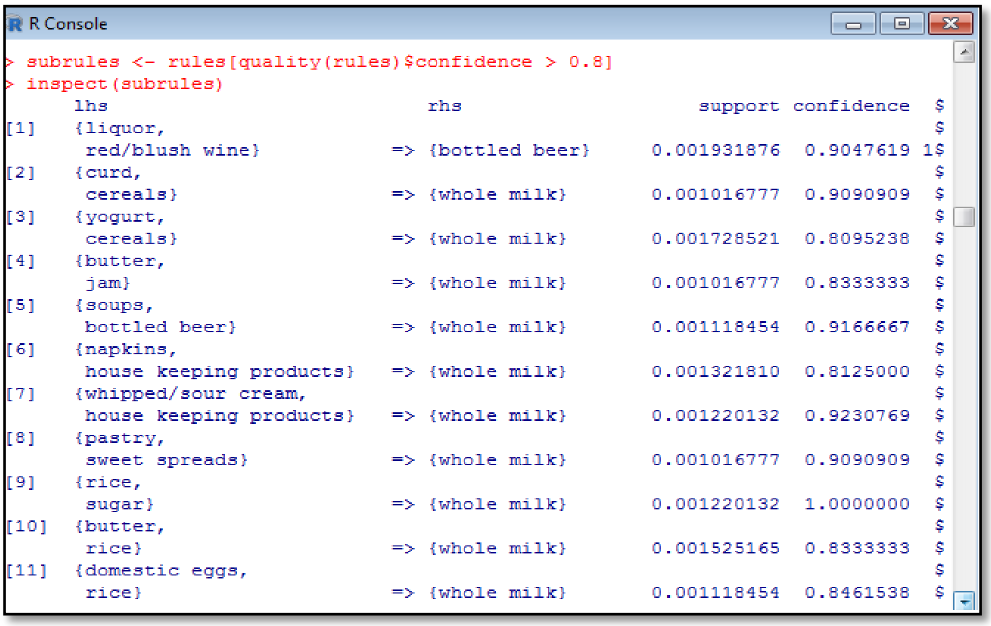


```
> confidence=0.5))
Error: unexpected ')' in "confidence=0.5)"
> rules <- apriori(Groceries, parameter=list(support=0.001,confidence=0.5))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
        0.5    0.1    1 none FALSE            TRUE       5   0.001      1
 maxlen target   ext
     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 9

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
sorting and recoding items ... [157 item(s)] done [0.02s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 done [0.01s].
writing ... [5668 rule(s)] done [0.00s].
creating S4 object  ... done [0.02s].
```

Fig 6: Mining Association rules

4.7 Extract the Rules in which the Confidence Value is >0.8 and high lift

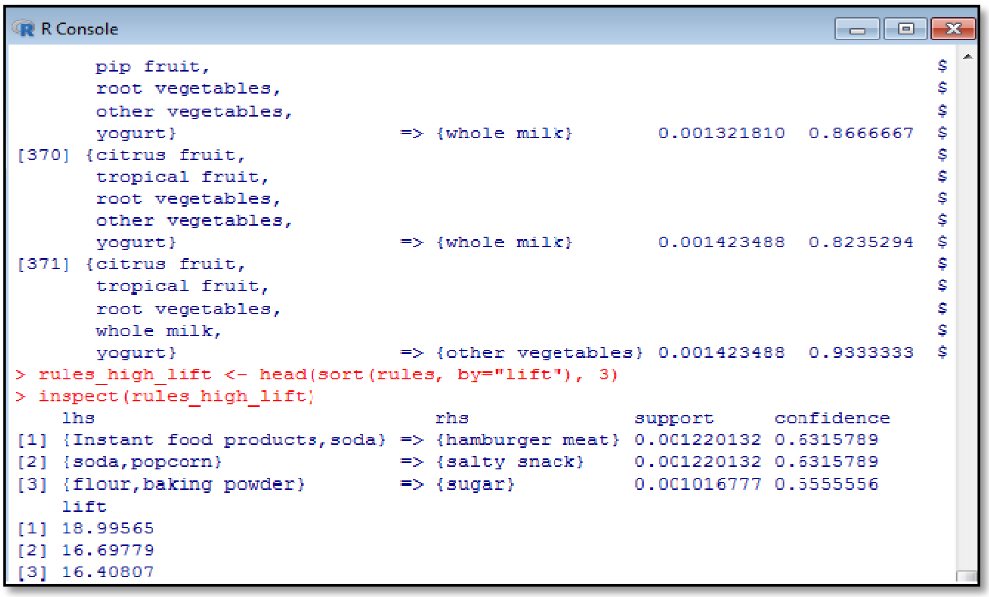
The rules for specific values of confidence and lift can be extracted by using the statements shown in figure 7.



```
> subrules <- rules[quality(rules)$confidence > 0.8]
> inspect(subrules)
      lhs                            rhs                   support  confidence $
[1]   {liquor,                                                                  $
       red/blush wine}            => {bottled beer}     0.001931876  0.9047619 1$
[2]   {curd,                                                                    $
       cereals}                   => {whole milk}       0.001016777  0.9090909 $
[3]   {yogurt,                                                                  $
       cereals}                   => {whole milk}       0.001728521  0.8095238 $
[4]   {butter,                                                                  $
       jam}                       => {whole milk}       0.001016777  0.8333333 $
[5]   {soups,                                                                   $
       bottled beer}              => {whole milk}       0.001118454  0.9166667 $
[6]   {napkins,                                                                 $
       house keeping products}    => {whole milk}       0.001321810  0.8125000 $
[7]   {whipped/sour cream,                                                      $
       house keeping products}    => {whole milk}       0.001220132  0.9230769 $
[8]   {pastry,                                                                  $
       sweet spreads}             => {whole milk}       0.001016777  0.9090909 $
[9]   {rice,                                                                    $
       sugar}                     => {whole milk}       0.001220132  1.0000000 $
[10]  {butter,                                                                  $
       rice}                      => {whole milk}       0.001525165  0.8333333 $
[11]  {domestic eggs,                                                           $
       rice}                      => {whole milk}       0.001118454  0.8461538 $
```

Fig 7: Rule extraction

Results can be reviewed by providing different values of confidence. The top three rules with high threshold can be extracted for the parameter "lift" using the statements described in figure 8:



```
        pip fruit,                                                              $
        root vegetables,                                                        $
        other vegetables,                                                       $
        yogurt}                   => {whole milk}       0.001321810  0.8666667  $
[370] {citrus fruit,                                                            $
        tropical fruit,                                                         $
        root vegetables,                                                        $
        other vegetables,                                                       $
        yogurt}                   => {whole milk}       0.001423488  0.8235294  $
[371] {citrus fruit,                                                            $
        tropical fruit,                                                         $
        root vegetables,                                                        $
        whole milk,                                                             $
        yogurt}                   => {other vegetables} 0.001423488  0.9333333  $
> rules_high_lift <- head(sort(rules, by="lift"), 3)
> inspect(rules_high_lift)
    lhs                            rhs                support     confidence
[1] {Instant food products,soda} => {hamburger meat} 0.001220132 0.6315789
[2] {soda,popcorn}               => {salty snack}    0.001220132 0.6315789
[3] {flour,baking powder}        => {sugar}          0.001016777 0.5555556
    lift
[1] 18.99565
[2] 16.69779
[3] 16.40807
```

Fig 8. Rules with high thresholds

5. CONCLUSION

Association rule mining is among the most prominent methodology in data mining, to determine the interesting relations among the data stored in large database.

R provides prominent and powerful tools for statistical data analysis and model building. Part of its power comes from the fact that users can contribute data mining capabilities to it in the form of external libraries or packages, which can then be used by anyone. Its vast number of community developers resulted in a immeasurable packages capable of handling a wide range of data mining techniques.

This paper presents the effectiveness of the "arules" package in mining the association rules based on different parameters. Apriori Algorithm is used to discover and understand the underlying patterns involved in the groceries dataset records from the data contents in various sections.

Future work includes using some approach with Apriori algorithm which has very less number of scans of database. Another solution to improve the efficiency is division of large database among processors.

6. REFERENCES

[1] Suriya, Shantharajah and Deepalakshmi 2012. A Complete Survey on Association Rule Mining with Relevance to Different Domain. International journal of advanced scientific and technical research,

[2] Zhu, Z., Wang, J. 2007. Book recommendation on service by improved association rule mining algorithms. In the

proceeding of international conference on Machine Learning and Cybernetics, pp. 3864-3869.

[3] Feng Yucai, 1998 .Association Rules Incremental Updating Algorithm, Journal of Software.

[4] Jaiwei Han and Micheline Kamber. Data Mining Concepts and Techniques, Second Edition, Morgan Kaufmann Publishers.

[5] Ms.Shweta and Dr. Kanwal Garg. 2013. Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms. International Journal of Advanced Research in Computer Science and Software Engineering ISSN: 2277 128X Volume 3, Issue 6.

[6] Lei Guoping, Dai Minlu, Tan Zefu and Wang Yan.2011. The Research of CMMB Wireless Network Analysis Based on Data Mining Association Rules. IEEE conference paper project supported by the Science and Technology Research Project of Chongqing municipal education commision under contract no KJ101114 and KJ 111103.

[7] Lin, H., Goumin ,Z., Liu, Q. 2009. Application of Apriori Algorithm to Data Mining of the Wildfire. In the proceeding of 6th International Conference on Fuzzy Systems and Knowledge Discovery.

[8] Agarwal, R. C., Aggarwal, C. C., and Prasad, V. V. V. 2000. Depth first generation of long patterns. In Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 108–118.

[9] Borgelt, C. 2003. Efficient implementations of apriori and eclat. In Goethals, B. and Zaki, M. J.,editors, Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations, Melbourne, FL, USA.

[10] Creighton, C. and Hanash, S. 2003. Mining gene expression databases for association rules. Bioinformatics, 19(1):79–86.

[11] DuMouchel and Pregibon, D.2001.Empirical bayes screening for multi-item associations. In Provost, F. and Srikant, R., editors, Proceedings of the 7th ACM SIGKDD Intentional Conference on Knowledge Discovery in Databases and Data Mining, pages 67–76. ACM Press.