JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

Computational Study on Association Rule Mining Using Microarray Data

K. Mohan Kumar[1], S. Devi[2]

[1,2]Dept. of Computer Science, Rajah Serfoji Govt. College, Thanjavur, India

*Corresponding Author: deviraman.72@gmail.com

Available online at: www.ijcseonline.org

Accepted: 22/Nov/2018, Published: 30/Nov/2018

***Abstract*—**Data mining is used to bring out the unknown information from known large data set. In data mining Association Rule Mining (ARM) is a technique which discovers the frequent relation between the patterns by using the terms support and confidence. Apriori, Partition, Border and Incremental algorithms are some of the algorithms in ARM. In this work microarray dataset for psychological disorders is extracted from GEO data base, applied Apriori algorithm, implemented using R tool and recognized the relationship between the diseases in psychological disorder.

Keywords: Association Rule Mining, Apriori, Microarray dataset, Psychological Disorder, Occurrences

I. INTRODUCTION

Data mining is a technique which helps to extract important data from a large database. It is a process of sorting through large amounts of data and picking out relevant information through the use of certain sophisticated algorithms. The amount of data accumulated is increased every piece of second. Thus, volume of data gathered in different situations is available for further reference. In this circumstance, data mining become an important tool to transform the gathered data into essential information. This evolution began when business data was first stored on computers, continued with improvements in data access. Recent developments in technologies allow users to store the data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Its functions include clustering, classification, prediction, and link analysis / associations. [1].

The rest of the paper is organized as follows. Section 1.1 to 1.4 explains the basic concepts of ARM, GEO database,R-Tool and Apriori algorithm respectively. Section II contains the methodology to find the relationship between the Psychological Disorder Occurrences, Section III discussed the implementation of Apriori in R, Section VI discussed the result and finally in the Section V concludes the paper and gives the future scope.

1.1 ASSOCIATION RULE MINING (ARM)

Association rule mining (ARM) is a most important data mining task that works in an unsupervised manner. It finds the rules that explore the relationships among the data items based on their occurrences in transactions. A traditional approach like Market Basket Analysis is used to find the frequent patterns but may not be applicable for several real life applications. ARM technique is used in some real time applications such as Web mining, scientific data analysis, Bioinformatics and Medical diagnosis. ARM is measured by using the terms support and confidence. It identifies the relationships and rules generated by analysing data for frequently used if/then patterns. Association rules are usually needed to satisfy a user-specified minimum support and a user-specified minimum confidence simultaneously. Some of the algorithms utilizing association rule mining (ARM) are Apriori, Partition, Border Algorithm and Incremental Algorithm [2].

ARM is the prominent model invented and extensively studied by databases and data mining community. Association mining has been used in many application domains. One of the best known is the business field where discovering of purchase patterns or association between products is very useful for decision making and effective marketing [3, 4].

Most of the association rules are generated by counting the number of occurrences of the rule in the database. Association rules are statements of the form (X1, ..Xn) -> Y which means that Y may be present in the transaction if X1, X2,…Xn are all the factors present in the transaction. The association rule will be useful for discovering interesting relationships hidden in the large data base. Support determines how often a rule is applicable to a given data set, while confidence determines how frequently Y appears in the transactions that contain X factors. Support is an important measure of the interestingness of the rule. The rule with very low support may occur simply by chance. Support is often used to eliminate uninteresting rules. Confidence on the other hand measures the reliability of the rule. There are different types of association rules classified under Boolean association rule and generalized association rule. The simplest form is the type that only shows valid or invalid association. This Boolean nature of rule gives the name Boolean Association Rules. Rules that are accumulating into several association rules together are called Multilevel or Generalized association rule [5].

1.2 GEO DATABASE

The Gene Expression Omnibus (GEO) project was initiated at NCBI in 1999 in response to the growing demand for a public repository for data generated from high-throughput gene expression microarray experiments. GEO provides a flexible and open design that facilitates submission, storage and retrieval of heterogeneous data sets from high–throughput gene expression and genomic hybridization experiments. GEO was never intended to replace lab-specific gene expression databases or laboratory information management systems (LIMS), both of which usually cater to a particular type of data set and analytical method. Rather, GEO complements these resources by acting as a central, molecular abundance–data distribution hub. The GEO resource is under constant development and aims to improve its indexing, linking searching, and display capabilities to allow vigorous data mining. Because the data sets stored within GEO are from heterogeneous techniques and sources, they are not necessarily comparable [6].

The three central data entities of GEO are platforms, samples and series and were designed with gene expression and genomic hybridization experiments in mind. A platform is essentially a list of probes that define what set of molecules may be detected. A sample describes the set of molecules that are being probed and references a single platform used to generate its molecular abundance data. A series organizes samples into the meaningful data sets which make up an experiment. The GEO repository is publicly accessible through the World Wide Web at **http://www.ncbi.nlm.nih.gov/geo**.There has been a great explosion of genomic data in recent years. This is due to the advances in various high-throughput biotechnologies such as gene expression microarrays. The GEO database stores molecular abundance data generated by a wide variety of microarray-based experiments that measure gene expression or detect genomic gains and losses. These large genomic data sets are information-rich and often contain much more information than the researchers who generated the data might have anticipated. Such an enormous data volume enables new types of analyses, but also makes it difficult to answer research questions using traditional methods. Gene expression data can be a valuable tool in understanding of genes, biological networks, and cellular states. Analysis of these massive genomic data helps to determine how the expression of any particular gene might affect the expression of other genes and also identifies what genes are expressed in diseased cells that are not expressed in healthy cells. Association rule mining widely used in the area of market basket analysis, can also be applied in the analysis of gene expression data as well. It can reveal biologically relevant associations between different genes or between environmental effects and gene expression [7].

1.3 R-TOOL

R is an open-source software environment for statistical computing and graphics. R compiles and runs on Windows, Mac OS X, and numerous UNIX platforms (such as Linux). For most platforms, R is distributed in binary format for ease of installation. This language was very much influenced by the S language, which was originally developed at Bell Laboratories by John Chambers and colleagues. With the base talents of R’s core development team, R has evolved into abridge language for statistical computations in many disciplines of academics and various industries is designed around its core scripting language but also allows integration with compiled code written in C, C++, Fortran, Java, etc., for computationally intensive tasks or for leveraging tools provided for other languages. R, like other programming languages, is extended (or developed) through user-written functions. An integrated development environment, such as R Studio, is designed to facilitate such work. In addition, unlike many other statistical software packages in which a graphical user interface is employed; a typical user interacts with R primarily through the command line. Like R, R Studio is an open-source project. Its stated goal is to develop a powerful tool that supports the practices and techniques required for creating trustworthy, high-quality analysis. In this study the R studio is used to manipulate the genomic data implementing Apriori algorithm type under association rule mining [8].

1.4 APRIORI ALGORITHM

Apriori algorithm is used for frequent item set mining and association rule learning. It is devised to operate on a database containing a lot of transactions, for instance, items brought by customers in a store. It has also been used in the field of healthcare for the detection of adverse drug reactions. It produces association rules that indicate what all combinations of medications and patient characteristics lead to definite results. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database [9].

II. METHODOLOGY

The dataset is taken from GEO online database and given in the following Table-1. The set of data items are Schizophrenia, Bipolar_ Disorder, insomnia, Alzheimer, Parkinson. It consists of 6 occurrences. Every occurrence is in an ordered row, denoted as 0's and 1's where 0 is the absence of an item and 1 is the presence of an item. In this example the rule might be { Schizophrenia, Bipolar_ Disorder } =>{insomnia}, which means that if a patient is affected by Schizophrenia or Bipolar_ Disorder, then there may be a possibility of Insomnia.

Table 1: Psychological Disorder Occurrences

OCCURRENCES	**INSOMNIA (INSO)**	**ALZHEIMER (ALZH)**	**BIPOLAR_ DISORDER (BP_D)**	**SCHIZOPHRENIA (SCHIZ)**	**PARKINSON (PARK)**
O1	1	0	1	1	0
O2	1	1	1	0	0
O3	0	1	0	0	1
O4	0	1	1	1	0
O5	1	0	1	1	1
O6	1	1	1	1	1

The above example is applied in Apriori with the support threshold set as 50%, i.e. the occurrences of the items are significant only if the support is more than 50%. The following Phases are used to apply the Appriori algorithm.

Phase 1: Create periodic occurrence of items using the data in Table-1.

Table 2: Periodic occurrence of Items

Item	Frequency
Schizophrenia (SCHIZ)	**4**
Bipolar_ Disorder (BP_D)	**5**
Insomnia(INSO)	**4**
Alzheimer(ALZH)	**4**
Parkinson(PARK)	3

Phase 2: Here, the support threshold is 50%. So, the items that occur more than three times are considered as significant. The Table-3 exhibits the items that have more than 3 occurrences. They are Schizophrenia, Bipolar_Disorder, Insomnia and Alzheimer.

Table 3: Periodic occurrence of items with significant

Item	Frequency
Schizophrenia (SCHIZ)	4
Bipolar_ Disorder (BP_D)	5
Insomnia(INSO)	4
Alzheimer(ALZH)	4

Phase 3: Create all the possible pairs of the potential items. In this pair creation assume pairs which are in reverse order are equal. For example XY is equal to YX. The pairs for first item are SCHIZ-BP_D,SCHIZ-INSO, SCHIZ-ALZH, SCHIZ-PARK.. Similarly the pair for the second, third, fourth items are created. The final pairs are SCHIZ-BP_D,SCHIZ-INSO, SCHIZ-ALZH, SCHIZ-PARK, BP_D-INSO, BP_D-ALZH, BP_D-PARK, INSO-ALZH, INSO-PARK, ALZH-PARK.

Phase 4: Count the appearances of each pair using Table -1 Table – 4 shows the result.

Table 4 :Periodic occurrences with 2 item sets

Item	Frequency
SCHIZ-BP_D	**4**
SCHIZ-INSO	**3**
SCHIZ-ALZH	2
SCHIZ-PARK	2
BP_D-INSO	**4**
BP_D-ALZH	**3**
BP_D-PARK	2
INSO-ALZH	2
INSO-PARK	2
ALZH-PARK	2

Phase 5: Item sets that are above the support threshold will be considered as significant. They are SCHIZ-BP_D, SCHIZ-INSO, BP_D-INSO and BP_D-ALZH

Phase 6: The probabilities of occurrences of three item sets are created with the help of item sets which are paired in Phase 5. SCHIZ-BP_D, SCHIZ-INSO which gives SCHIZ-BP_D-INSO and BP_D-INSO, BP_D-ALZH which gives BP_D-INSO-ALZH. The following Table-5 shows the occurrences of these two item sets.

Table 5: Periodic occurrences with 3-itemset

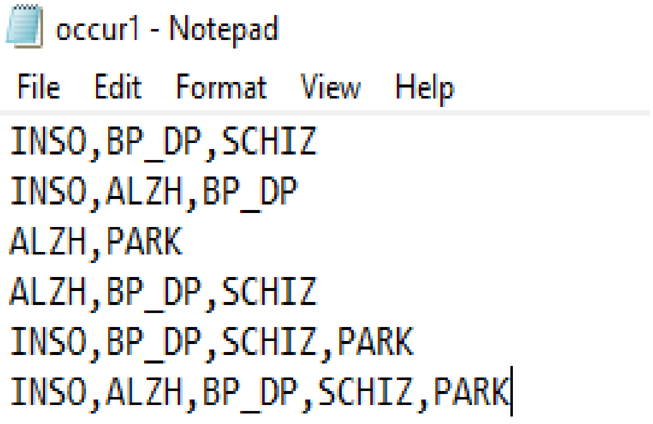
Item	Frequency
SCHIZ-BP_D-INSO	3
BP_D-INSO-ALZH	2

The above Table-5 shows that SCHIZ-BP_D-INSO is the only significant item set with low support threshold. Hence, the three set items SCHIZ-BP_D-INSO is the most occurring item i.e., the psychological disorders Schizophrenia and Bipolar_Disorder has the possibility of occurring Insomnia.

III. IMPLEMENTATION OF APRIORI IN R

Before implementing the apriori algorithm a package called *"arules"* should be installed.

A text file is created to apply the above example of psychological disorder occurrences.This file contains the saved item sets of occurrences as shown in the Figure1.



```
INSO,BP_DP,SCHIZ
INSO,ALZH,BP_DP
ALZH,PARK
ALZH,BP_DP,SCHIZ
INSO,BP_DP,SCHIZ,PARK
INSO,ALZH,BP_DP,SCHIZ,PARK
```

Figure 1: File containoccurrences of the item sets

The "read" command is used to read the input file. The support and confidence to determine the association rule is evaluated and obtained as support->0.5 and confidence->0.7. With the help of image command the occurrences of the item sets are displayed in rows and columns as shown in the Figure2.

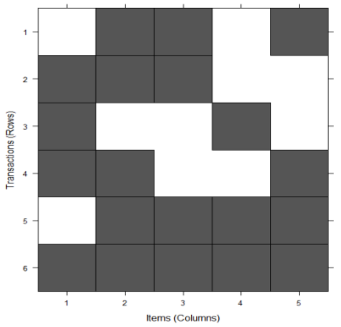
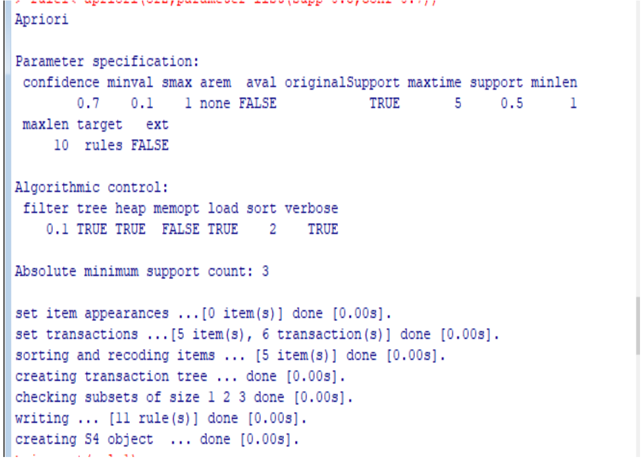


Figure 2: Image shows occurrences of the item sets

To run Apriori algorithm in R, use the command Apriori along with the parameters such as the file which consists of psychological disorders occurrences data base, support of 0.5 and confidence of 0.7. The output is shown in the Figure - 3.

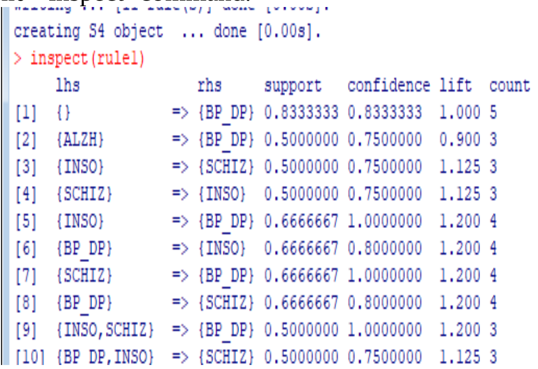


```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
        0.7    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 ...[5 item(s), 6 transaction(s)] done [0.00s].
sorting and recoding items ... [5 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 done [0.00s].
writing ... [11 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

Figure – 3: Result of R for the given input

IV. RESULTS AND DISCUSSIONS

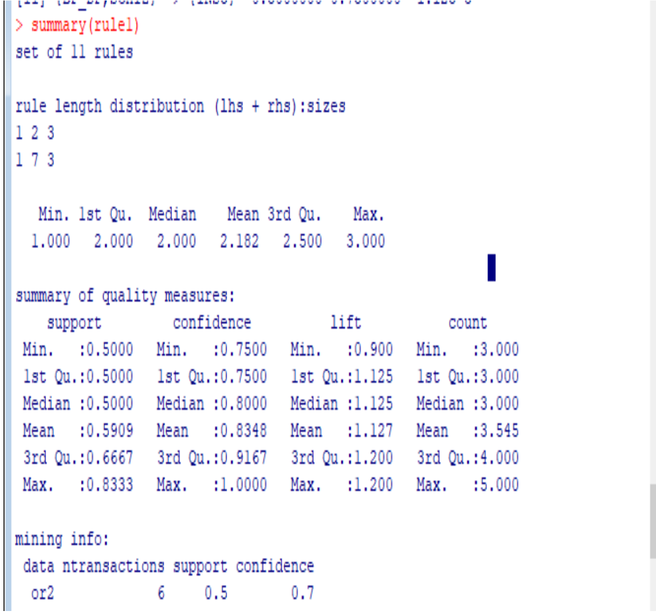
The outputs of the Apriori algorithm can be inspected by using the ' inspect' command.



```
creating S4 object  ... done [0.00s].
> inspect(rule1)
     lhs              rhs     support   confidence lift  count
[1]  {}            => {BP_DP} 0.8333333 0.8333333  1.000 5
[2]  {ALZH}        => {BP_DP} 0.5000000 0.7500000  0.900 3
[3]  {INSO}        => {SCHIZ} 0.5000000 0.7500000  1.125 3
[4]  {SCHIZ}       => {INSO}  0.5000000 0.7500000  1.125 3
[5]  {INSO}        => {BP_DP} 0.6666667 1.0000000  1.200 4
[6]  {BP_DP}       => {INSO}  0.6666667 0.8000000  1.200 4
[7]  {SCHIZ}       => {BP_DP} 0.6666667 1.0000000  1.200 4
[8]  {BP_DP}       => {SCHIZ} 0.6666667 0.8000000  1.200 4
[9]  {INSO,SCHIZ}  => {BP_DP} 0.5000000 1.0000000  1.200 3
[10] {BP_DP,INSO}  => {SCHIZ} 0.5000000 0.7500000  1.125 3
```

Fig 4

The result infers that the patients affected with schizophrenia may also be affected with bipolar disorder which has a greater lift of 1.20 and count of 4 occurrences. The same way the person affected with bipolar disorder has the possibility to be affected with insomnia which has 80% confidence, lift of 1.2 and count of 4 occurrences. Finally, the persons who are affected with schizophrenia and insomnia have the possibility of getting bipolar disorder because of its 100% reliability (1.0 confidence) and the greater lift of 1.2 with 3 occurrences. Here, the support extends to the maximum of 80% and the confidence extends up to 100%. Therefore, the rules are interestingly associated. The summary is given below in the Figure 5.



```
> summary(rule1)
set of 11 rules

rule length distribution (lhs + rhs):sizes
1 2 3
1 7 3

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   2.000   2.182   2.500   3.000

summary of quality measures:
    support          confidence          lift            count
 Min.   :0.5000   Min.   :0.7500   Min.   :0.900   Min.   :3.000
 1st Qu.:0.5000   1st Qu.:0.7500   1st Qu.:1.125   1st Qu.:3.000
 Median :0.5000   Median :0.8000   Median :1.125   Median :3.000
 Mean   :0.5909   Mean   :0.8348   Mean   :1.127   Mean   :3.545
 3rd Qu.:0.6667   3rd Qu.:0.9167   3rd Qu.:1.200   3rd Qu.:4.000
 Max.   :0.8333   Max.   :1.0000   Max.   :1.200   Max.   :5.000

mining info:
 data ntransactions support confidence
  or2             6     0.5        0.7
```

Figure-5

The summary interprets the association between schizophrenia, bipolar disorder and insomnia which have interesting relationship. This has been proved by the support of 50% chances of occurrences and the confidence 70% that denotes the reliability of the rule. Hence, the patients affected with schizophrenia and bipolar disorder, have a maximum possibility of being affected with insomnia.

V. CONCLUSION

This paper describes how the association rule should be verified manually for the given data set and analysed using R tool. The association between schizophrenia, bipolar disorder and insomnia are identified using association rule mining and studied the implementation of R tool with psychological disorder data set. Diseases and their relationships for a patient is a most wanted finding because it helps to find the subsequence of one problem to another. This method can be applied for larger medical data sets which are used to predict the relationships among the diseases.

REFERENCES

[1]. Doddi S, Marathe A, Ravi SS, and Torney DC, "Discovery of association rules in medical data.", Med. Inform. Internet. Med.,vol. 26 (2001): 25–33.

[2]. Arockiam L, Baskar SS, and Jeyasimman L, "Importance of Association Rules in Data Mining: A Review, International Journal of Soft Computing." 7(3), (2012): 135-140.

[3]. Tuzhilin A, Adomavicius G, Zatane O, Goebel R, Hand D, Keim D, NgR, "Handling very large numbers of association rules in the analysis of microarray data", Proc. of the 8th ACM SIGKDD international conference on knowledge discovery and data mining, (2002): 396-404.

[4]. Anandhavalli M, Ghose MK, and Gauthaman K, "Association Rule Mining in Genomics." International Journal of Computer Theory and Engineering, Vol. 2. No. 2 (2010):1793-8201.

[5]. RakeshAgrawal and RamakrishnanSrikant, "Mining Sequential Patterns." In Proc. of the 11th International Conference on Data Engineering, Taipei, Taiwan, (1995).

[6]. Tanya Barrett and Ron Edgar, "Mining Microarray Data at NCBI's Gene Expression Omnibus (GEO)." National Institute of Health Public Access (NIH PA), Methods Mol Biol. (2006); 338: 175–190.

[7]. Ron Edgar and Alex Lash, "The Gene Expression Omnibus (GEO): A Gene Expression and Hybridization Repository." The NCBI Handbook (2015): 6-17, https://www.researchgate.net.

[8]. Rashmi Jain, "Machine Learning", (2017). Available at https://www.hackerearth.com/blog/machine-learning/beginners-tutorial-apriori-algorithm-data-mining-r-implementation/

[9]. Trupti A Kumbhare and Santosh V Chobe, "An Overview of Association Rule Mining Algorithms." International Journal of Computer Science and Information Technologies, Vol. 5 (1) (2014): 927-930.

Authors Profile

Dr.K.Mohan Kumar received Master of Computer Science, Ph.D in Computer Science from Bharathidasan University, Tiruchirappalli, India and M.Phil computer science from Manonmaniyam Sundaranar University, Thirunelveli, India. He is currently working in PG and Research Department of Computer Science, Rajah Serfoji Government College, Thanjavur, T.N, India. His main research work focuses on IoT, Cloud computing, Network Security, Big Data Analytics and Computational Intelligence based education. He has published more than 50 research papers in reputed International journals. He has 23 years of teaching experience and18 years of Research experience.

S. Devi received Master of Computer Application from Anna University, Chennai, India and M.Phil. computer science from PRIST University, Vallam, Thanjavur, India. She is currently working in SRM University,Ramapuram,Chennai, T.N, India. Her main research work focuses on Data Mining, Cloud computing, Network Security, Big Data Analytics and Computational Intelligence based education. She has published more than 2 research papers in reputed International journals. She has 5 years of teaching experience.