Association Analysis of Bank Dataset Utilizing Apriori Algorithm in Weka and RStudio

Arvy Olarcos Llave
College of Science
Bicol University
arvyolarcos.llave@bicol-u.edu.ph

Alliana Figueras Ermino
College of Science
Bicol University
allianafigueras.ermino@bicol-u.edu.ph

Ella Mae Ronquillo Poche
College of Science
Bicol University
ellamaeronquillo.poche@bicol-u.edu.ph

ABSTRACT

Nowadays, a bank is one of the most common institutions that provides and handles cash, credit, and other financial transactions for individual consumers and businesses. It promotes the safety and security of an individual in terms of handling money. By that, association rules in data mining are helpful in supporting the systems in collecting and associating the information given by the customer. This study aims to perform association analysis utilizing apriori algorithm in Weka and RStudio. Whereas in R programming, it shows the structure of the transaction checked out the most frequent items. Moreover, it also creates model creation that finds the most frequent itemsets (target = "frequent") with the default settings and is sorted by support. After that, we can also create a model using apriori in which the main objective is to show the rules and look for the highest lift. On the other hand, performing apriori in Weka shows the result based on confidence and based on the lift.

Keywords

Bank; Association Analysis; Apriori Algorithm; RStudio; Weka; Data Mining

1. INTRODUCTION

In this world full of data, data mining is essential in collecting and transforming raw data into useful information. Data mining is used in an enormous amount of data to understand and discern patterns and trends. It is beneficial to any large company in the world.

One of the most used data mining types is association rule mining. Association rules are "if-then" statements that show the probability and relationships of data items. According to Saxena and Rajpoot [1], association rule mining is concerned with knowledge. Correlation concepts are utilized in a variety of fields to identify data patterns. Using trends, we can see how many different types of occurrences occur at the same time.

Apriori algorithm is a technique in association rule. It is a collection of actions to be taken to determine the most frequent itemset in a given database. This data mining approach iteratively applies the join and prune processes until the most frequent itemset is obtained. The problem specifies or the user assumes a minimum support threshold.

2. RELATED LITERATURE AND STUDIES

During economic and financial globalization, financial activities have increased and developed on a worldwide scale. As a result, the modern financial system's environment has become increasingly complicated, presenting a variety of complex forms [2]. As banks have become one of the most vital components of any financial system, ensuring the stability of the banking sector has gained significant importance as a policy initiative worldwide. Banking stability as an economic indicator can be used to determine whether an economy is robust enough to withstand both internal and external shocks. Banking stability is a function of several health parameters of individual banks [3].

Jisha and Kumar provided in their study the application areas of data mining techniques in banking. These are the following:

Security and fraud detection: Big secondary data like transaction records are monitored and analyzed to enhance banking security and distinguish the unusual behavior and patterns indicating fraud, phishing, or money laundering.

Risk management and investment banking: Analysis of in-house credit card data freely accessible for banks enables credit scoring and credit-granting which form part of the most popular tools for risk management and investment evaluation.

CRM: DM techniques have been widely applied in banking for marketing and customer relationship management-related purposes such as customer profiling, customer segmentation, and cross/up-selling. These help the banking sector to have a better understanding of their customers, predict customer behavior, accurately target potential customers, and further improve customer satisfaction with a strategic service design.

Other advanced supports: A few fewer mainstream applications focus on branching strategy, and efficiency and performance evaluation, which can significantly assist in achieving strategic branch locating and expansion plans [4].

According to Verma and Pathak, assurance of association rules regarding banking data is a challenging though demanding task because: (1) the volume of banking data is immense so, the data must be appropriately prepared before the final step of the application of the method; (2) the objective must be clearly stated from the beginning; and (3) good knowledge of the dataset is necessary not only for the data miner but also for the final analyst, if not false results will be produced by the data miner and inaccurate conclusions will be made by the manager [5].

Hasheminejad and Khorrami stated that identifying clients and analyzing their behavior is critical in today's corporate environment, especially for the banking industry. Customer Relationship Management (CRM) is the practice of sustaining lucrative customer relationships by providing value and loyalty to customers. Furthermore, CRM aids in the improvement of company interactions with clients. CRM's purpose is to optimize a customer's lifetime worth to an enterprise. Customer Lifetime Worth (CLV) may be used to rank and classify customers based on their lifetime value to find and keep valued customers. There

are numerous models for estimating CLV based on customer history data. This topic assists businesses in their efforts to retain important clients. Data mining techniques play a critical role in revealing buried knowledge and information. The purpose of their study is to examine data mining techniques used to analyze bank customers to assist banks in better identifying their clients and designing more efficient marketing tactics [6].

Aksu and Doğan used Weka program in their study. They aim to introduce one of the data mining methods which is very popular in recent years and commonly used in this area. In today's age of technology, the amount of information at hand is constantly increasing and the derivation of meaningful results from this information is seen as a valuable field of study [7].

3. MATERIALS AND METHODS

3.1 Dataset Description

The Bank Dataset is used to perform Association Analysis using R Programming and WEKA. Whereas it contains 187 instances and 23 attributes. The dataset contains 12 attributes and 600 instances in total.

Table 1. Bank Dataset

id	a unique identification number
age	age of customer in years (numeric)
sex	MALE / FEMALE
region	inner_city/rural/suburban/town
income	income of customer (numeric)
married	is the customer married (YES/NO
children	number of children (numeric)
car	does the customer own a car (YES/NO)
save_acct	does the customer have a saving account (YES/NO)
current_acct	does the customer have a current account (YES/NO)
mortgage	does the customer have a mortgage (YES/NO)
pep	did the customer buy a PEP (Personal Equity Plan) after the last mailing (YES/NO)

3.2 Weka Tool

In this experiment, the researchers used Waikato Environment for Knowledge Analysis or Weka, an open-source data mining tool. The University of Waikato in New Zealand created Weka, which uses the Java programming language to perform data mining techniques. Weka is a cutting-edge tool for creating machine learning (ML) techniques and applying them to real-world data mining problems. It is a collection of machine learning algorithms designed for data mining tasks. The methods are immediately applied to a dataset. Weka implements data preprocessing, feature reduction, classification, regression, clustering, and association rules algorithms. It also provides visualization tools. It may be used to implement new machine learning algorithms as well as to expand current techniques.

3.3 R Programming Tool

The researchers also used RStudio in this case study. RStudio is a free, open-source IDE (integrated development environment) for R, a programming language for statistical computing and graphics. It features a console, syntax-highlighting editor with direct code execution, charting, history, debugging, and workspace management capabilities. It is mainly used for data science, scientific research, and technical communication. It allows users to edit and analyze data in a variety of ways, including developing machine learning models and storing and displaying data.

3.4 Apriori Algorithm

The Apriori algorithm is also used as a programming tool as the material of this study. The Apriori algorithm employs a repeated method known as level-wise search, in which n-itemset is used to find (n + 1) -itemset. A 1-itemset is searched in the first phase by tracing data in the database to determine the number of occurrences of each item. Then compute the support value and collect items that satisfy the support value. The following formula may be used to calculate support values:

$$Support\ (A) = \frac{\Sigma \text{ transactions containing A}}{Total\ transactions}$$

L1 denotes items that meet the specified minimum support. Furthermore, L1 is used to do a join process or a combination of existing itemsets in order to find a 2-itemset or L2. After L2 is formed, L2 is used to form L3 (3-itemset) and until the process stops when there is no itemset that meets the support value. After finding a collection of itemset that meets the support value, the association rule is established that meets the minimum confidence set, by calculating the confidence value of each itemset, by the formula:

$$Confidence\ (A \Rightarrow B) = P(B|A) = \frac{\Sigma \text{ transactions containing A \& B}}{\Sigma \text{ transactions containing A}}$$

There are two main processes in apriori algorithm, namely:
a. Join. This process is done by combining items with other items until no combination can be formed.
b. Pruning. The pruning process is the result of the items that have been combined and then trimmed using the minimum support specified by the user

3.5 Data Preparation in Weka

The unprocessed data set has 12 attributes and 600 instances in total. The dataset preprocessing steps are done:

- The id attribute has been removed, as it is of no importance.
- Then discretization is done:
  - Under this first the children attribute is discretized to values 0,1,2 and 3 as this attribute takes only these 4 values.
  - The age and income attributes are also discretized into 3 bins each.

Below is a visualization of the attribute data after the pre-processing of the dataset.

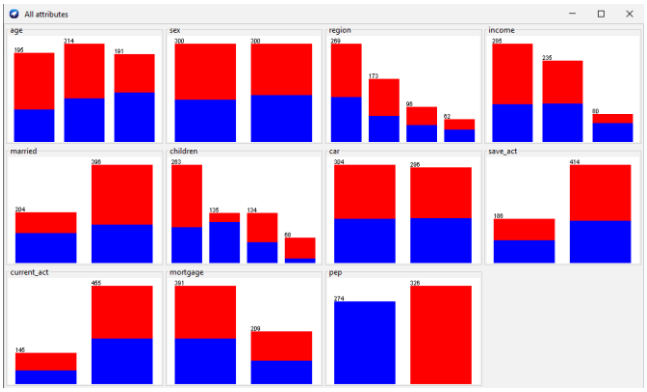

Figure 1. Attribute data after the pre-processing of the dataset.

3.6 Data Preparation in RStudio

In interpreting the association rule in R programming, apriori algorithm was used in this study. The dataset is composed of 600 objects,12 variables, and 1271 itemsets.

First, the required libraries are loaded in RStudio, and the .csv file of the bank dataset is imported. Then the dataset is converted into a set of transactions where each row represents a transaction and each column is translated into items. This is done using the constructor transactions(). The preparation of the data is all done using the arules package and then introduces the unified interface provided by arulesViz for association rule visualization.

The graph shows the inspection of the transactions. Checking the most frequent items in the dataset.

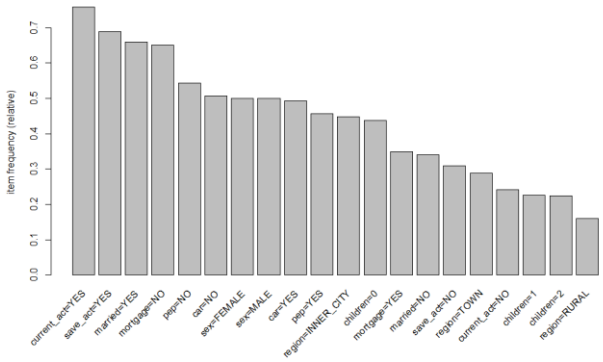

Figure 2. Inspection of the transaction.

4. RESULTS AND DISCUSSION

4.1 Weka Results

The class association rules generated by apriori algorithm on the pre-processed dataset are shown below. Results are derived from the properties which are described in Figure 3.

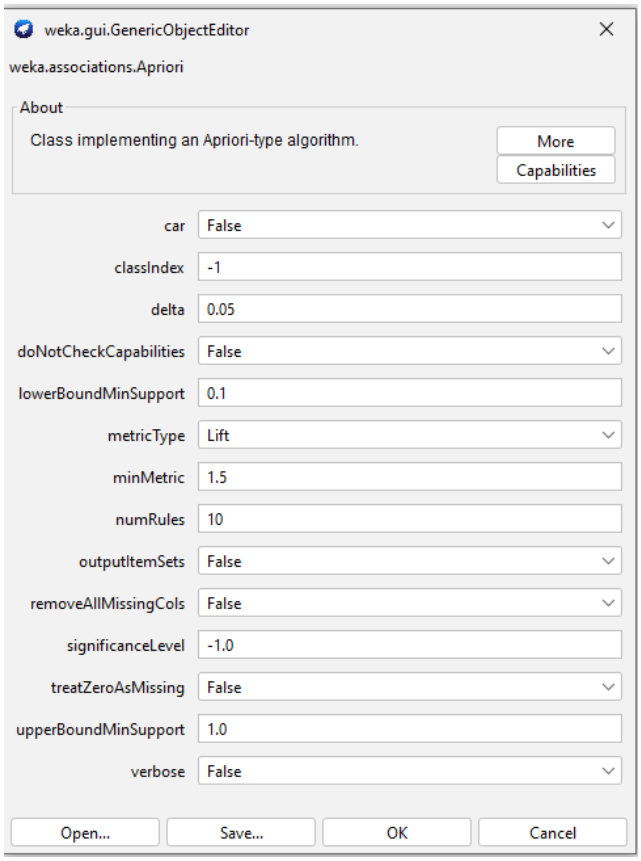

Figure 3. Shows the parameters using lift.

```
Apriori
=======

Minimum support: 0.15 (90 instances)
Minimum metric <lift>: 1.5
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 25

Size of set of large itemsets L(2): 166

Size of set of large itemsets L(3): 188

Size of set of large itemsets L(4): 38

Size of set of large itemsets L(5): 1

Best rules found:

 1. age=0_34 195 ==> income=0_24386 current_act=YES 138    conf:(0.71) < lift:(1.97)> lev:(0.11) [68] conv:(2.16)
 2. income=0_24386 current_act=YES 215 ==> age=0_34 138    conf:(0.64) < lift:(1.97)> lev:(0.11) [68] conv:(1.86)
 3. age=0_34 car=NO 107 ==> income=0_24386 100    conf:(0.93) < lift:(1.97)> lev:(0.08) [49] conv:(7.02)
 4. income=0_24386 285 ==> age=0_34 car=NO 100    conf:(0.35) < lift:(1.97)> lev:(0.08) [49] conv:(1.26)
 5. income=0_24386 pep=NO 176 ==> age=0_34 111    conf:(0.63) < lift:(1.94)> lev:(0.09) [53] conv:(1.8)
 6. age=0_34 195 ==> income=0_24386 pep=NO 111    conf:(0.57) < lift:(1.94)> lev:(0.09) [53] conv:(1.62)
 7. age=0_34 195 ==> income=0_24386 save_act=YES 106    conf:(0.54) < lift:(1.91)> lev:(0.08) [50] conv:(1.55)
 8. income=0_24386 save_act=YES 171 ==> age=0_34 106    conf:(0.62) < lift:(1.91)> lev:(0.08) [50] conv:(1.75)
 9. income=0_24386 285 ==> age=0_34 mortgage=NO 113    conf:(0.4) < lift:(1.9)> lev:(0.09) [53] conv:(1.3)
10. age=0_34 mortgage=NO 125 ==> income=0_24386 113    conf:(0.9) < lift:(1.9)> lev:(0.09) [53] conv:(5.05)
```

Figure 4. Results of associating rules based on lift.

4.2 RStudio Results

The class association rules were generated using apriori algorithm. The result below shows the top 10 rules generated based on lift.

```
> inspect(head(rules, n = 10))
     lhs                                                                rhs          support    confidence coverage   lift     count
[1]  {married=NO, children=0, save_act=YES, mortgage=NO}                -> {pep=YES} 0.05166667 0.9687500  0.05333333 2.121150 31
[2]  {married=NO, children=0, current_act=YES, mortgage=NO}             -> {pep=YES} 0.05833333 0.9459459  0.06166667 2.071414 35
[3]  {married=NO, children=0, mortgage=NO}                              -> {pep=YES} 0.07500000 0.9375000  0.08000000 2.052920 45
[4]  {children=0, save_act=NO, mortgage=YES}                            -> {pep=YES} 0.05666667 0.9189189  0.06166667 2.012231 34
[5]  {married=YES, children=1, save_act=YES, current_act=YES}           -> {pep=YES} 0.07333333 0.9166667  0.08000000 2.007299 44
[6]  {sex=FEMALE, married=YES, children=0, save_act=YES, current_act=YES, mortgage=NO} -> {pep=NO}  0.05833333 0.9459459  0.06166667 1.741005 35
[7]  {sex=FEMALE, married=YES, children=0, save_act=YES, current_act=YES} -> {pep=NO}  0.07333333 0.9361702  0.07833333 1.723013 44
[8]  {children=0, save_act=YES, current_act=YES, mortgage=YES}          -> {pep=NO}  0.07166667 0.9347826  0.07666667 1.720459 43
[9]  {married=YES, children=0, save_act=YES, current_act=YES, mortgage=NO} -> {pep=NO}  0.09500000 0.9344262  0.10166667 1.719803 57
[10] {region=TOWN, married=YES, children=0, save_act=YES}               -> {pep=NO}  0.06166667 0.9250000  0.06666667 1.702454 37
```

Figure 5. Top 10 rules generated based on lift.

Shown below are some visualizations of the rules in R.

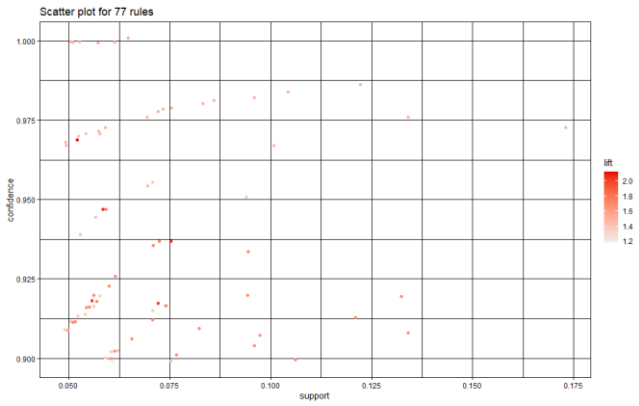

Figure 6. Scatter plot.

A scatter plot is a straightforward visualization of association rules to use a scatter plot with two interest measures on the axes. The default method for association rules in arulesViz is a scatter plot using support and confidence. A third measure (default: lift) is used as the color (gray level) of the points and the color key is provided to the right of the plot.

The stronger the correlation, the closer the data points are to forming a line. In this scatter plot, most of the rules have 0 to 0.100 support and there are groups of rules that are increasing. The trend of the rules seems divided according to the confidence. This shows that there is a trend in the data but has a weak correlation.

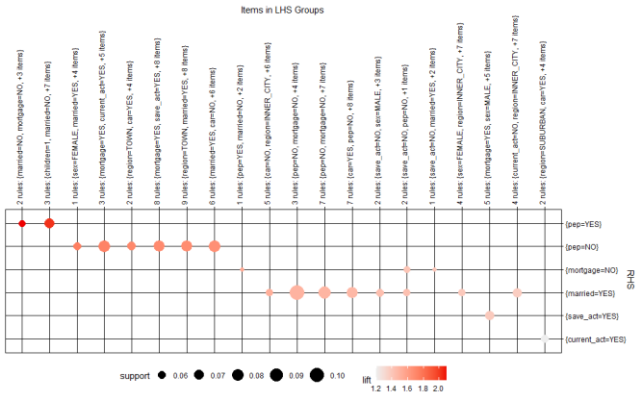

Figure 7. Graph-based visualization.

Graph-based visualization techniques visualize association rules using vertices and edges where vertices annotated with item labels represent items, and itemsets or rules are represented as a second set of vertices. Items relate to itemsets/rules using arrows. For rules arrows pointing from items to rule, vertices indicate LHS items and an arrow from a rule to an item indicates the RHS. Interest measures are typically added to the plot by using color or size of the vertices representing the itemsets/rules.

In this graph, the transaction pep = YES has the highest support of 0.07 and has a lift of 2.0 with 3 rules and 7 items. The transaction pep = NO, on the other hand, has the highest support of 0.08 and has a lift of around 1.6 with 6 rules and 6 items. The other transactions seem to have lower lifts and support according to the result.

5. CONCLUSION

At its most basic, association rule mining is using machine learning models to evaluate data in a database for patterns, or co-occurrences. This study focuses on understanding transactions using association rules in the bank dataset. The association rule used in this study is the apriori algorithm that runs on a massive database of transactions. After performing an apriori algorithm utilizing Weka and RStudio, it shows that in terms of flexibility, Weka is more convenient in performing association rules compared to RStudio. For a reason that Weka summarizes the process in which it provides a list of algorithms that can easily be accessed.

6. REFERENCES

[1] Saxena, A., & Rajpoot, V. (2021). A comparative analysis of association rule mining algorithms. IOP Conference Series: Materials Science and Engineering, 1099(1), 012032. https://doi.org/10.1088/1757-899x/1099/1/012032

[2] Gao, Q., Fan, H., & Shen, J. (2018). The stability of banking system based on network structure: An overview. *Journal of Mathematical Finance*, *08*(03), 517–526. https://doi.org/10.4236/jmf.2018.83032

[3] Al-Homaidi, E. A., Tabash, M. I., Farhan, N. H., & Almaqtari, F. A. (2019). The determinants of liquidity of Indian listed commercial banks: A panel data approach. Cogent Economics & Finance, 7(1), 1616521. https://doi.org/10.1080/23322039.2019.1616521

[4] Jisha, M., & Kumar, D. V. (2018). A case study on data mining applications on banking sector. International Journal of Computer Sciences and Engineering, 06(08), 67–70. https://doi.org/10.26438/ijcse/v6si8.6770

[5] Verma, N. N., & Pathak, D. (2019, April). A brief study of various data mining techniques and its applications in internet banking. Journal of Innovative Engineering and Research. Retrieved May 24, 2022, from https://jier.co.in/download/v2i1/1.Nidhi%20Nigam%20Verma%20pp%201-4.pdf

[6] Hasheminejad, S. M. H., & Khorrami, M. (2018, December 12). Data mining techniques for analyzing bank customers: A survey. IOS Press Content Library. Retrieved May 24, 2022, from https://content.iospress.com/articles/intelligent-decision-technologies/idt180335

[7] Aksu, G., & Doğan, N. (2019). An Analysis Program Used in Data Mining: WEKA. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 81–96. https://doi.org/10.21031/epod.399832