International Joint Conference - ICIEOM-ADINGOR-IISE-AIM-ASEM (IJC 2017)
Valencia, Spain, July 6-7, 2017

Data Mining Association Rules Applied to Supermarket Transactional Data Modeling: a case study in Brazil

Schonhorst G B[1], Paes V C[1], Balestrassi P P[1*], Paiva A P[1], Campos P H S[1]

Abstract The discovery of association rules is a data mining task that has been studied since the early 1990s. A major application of these rules is the Market Basket Analysis (MBA). In this type of problem the goal is to search for patterns in consumer behavior, acquiring knowledge of what products are usually brought together in a single purchase. The knowledge can be used to support decision making on operational and strategic levels. The growing interest of researchers in this area is due to both practical use and the difficulties and limitations present in this type of analysis. Yet real applications are still few. The objective of this paper is to conduct a market basket analysis through association rules mining on transactional data from a typical supermarket and hold a discussion on the applicability of the technique. The technique used in this study proved capable of generating large amount of useful knowledge for decision making. The definition of a specific focus proved to be crucial to the success of the analysis.

Keywords: market basket analysis, association rules, data mining;

1 Introduction

During years, predominantly manual methods had been used to transform data into knowledge. With large datasets, these methods were dispendious, both in financial terms and time consumption. The analysis was subjective and in most of the time impracticable (Fayyad et al., 1996). The search for efficient and faster methods for acquiring knowledge from the dataset stimulated the research in the area, that is

[*]Corresponding author: Pedro Paulo Balestrassi (e-mail: ppbalestrassi@gmail.com)
[1]Industrial Engineering and Management Institute. Federal University of Itajubá. Av. BPS 1303, Minas Gerais, Brazil

now know in the literature as Knowledge Discovery in Database (KDD) and Data Mining.

The discovery of association rules is a technique of data mining that is being studied since the beginning of the decade of 1990. The association task look to characterize how much the presence of a set of items in the registers of a database implies in the presence of some another distinct set of items in the same register (Agrawal and Srikant, 1994). According to Piatetsky-Shapiro (2007), great success was achieved in the performance improvement of the association fetching algorithms rules and many of them had been widely accepted, but applications in real problems are still few.

One of the main applications of the association rules is in the Market Basket Analysis (MBA). In this kind of problem the goal is to search behavior patterns of the consumers, acquiring knowledge, for example, of which products they choose to be led together in the same purchase (Silverstein et al., 1998). The knowledge acquired through an analysis of the consumer basket purchase can be used to support decisions in operation and strategical levels. According to Chen et al. (2005), the MBA can help the organization managers in the layout project, e-commerce, mix of products and other marketing strategies.

Analytical tools for decision aid are helping retail to conquer new customers and to keep loyalty those that already they possess. According to Grewal and Levy (2007), the utilization of data analysis techniques for the retail decision aid related to the CRM (Customer Relationship Management) in general and for rewards programs in particular represent new and promising areas for academic research.

The objective of this article is to carry through an analysis of purchase basket through association rules mining on transactional data from a typical Brazillian supermarket and to carry through a quarrel on the applicability of this type of task of data mining. In Section 2 will be explained the involved concepts in the process of data mining and the concept related to the association rules. In Section 3 will be displayed the process on how the data from the market basket was obtained and processed to acquire the results. The applicability of the technique is also argued. Finally, in the Section 4, the final conclusions and considerations are described.

2 Theoretical Foundation

2.1 Association Rules Mining and MBA

Through the last 10 years the data mining evolved, having been influenced for external forces, such as the growth of the electronic commerce and great progress in molecular biology. New research are appeared, such web mining, and open source

software were developed, as the WEKA, what led to spread of data mining applicability (Piatetsky-Shapiro, 2007).

2.2 Knowledge Discovery in Database (KDD)

Fayyad et al. (1996) defines the process of knowledge discovery based on data being the identification process of valid patterns, new, and potentially useful and understandable. The KDD Stages can be summarized, according to the authors:

- Selection: database attainment - or variables subgroup and samples on which the discovery process will be applied.
- Preprocessing: accomplishment of basic operations, as noise removal and definition of strategies to deal with missing, incomplete or inconsistent data.
- Transformation: possibly necessary operations: summary, generalization, normalization and attribute creation. The dimension reduction can be applied or another method to reduce the number of variables.
- Data Mining: data mining algorithm application on the dataset.
- Interpretation/Evaluation: evaluation of the results by a specialist in the area, or by knowledge measures already in the literature.

2.3 Association Rules

The discovery of association rules is a data mining technique that has received great attention from researchers. According to Hipp et al. (2002), the association has become a very popular mining technique due to its applicability to business problems with its inherent understandability, because even not experts in data mining can understand them.

An association rule is represented as an implication in the form of LHS → RHS, so that LHS and RHS are respectively the previous (Left Hand Side) and the resulting (Right Hand Side) rule. Association rules were defined by Agrawal and Srikant (1994) as:

"D is a database consisting of a set of items A = {a1, ..., an} ordered lexicographically, and a set of transactions T = $\{t_1, ..., t_n\}$, where each transaction ti $\in$ T is composed of a set of items (called itemsets) such that $t_i \subseteq A$. The transaction t_i supports the itemset X if $X \subseteq t_i$. The support P(X) of an itemset X represents the probability of occurrence of event X.

The association rule is an implication in the form LHS → RHS, in which LHS $\subset$ A, RHS $\subset$ A e LHS $\cap$ RHS = Ø. The rule LHS → RHS occurs in the set of transactions T with confidence conf and support sup, where P(LHS, RHS) represents the

support rule (the likelihood of the transaction LHS ∪ RHS) and P(RHS | LHS) the confidence of the rule (the conditional probability RHS given LHS)".

2.4 Market Basket Analysis

A major application of association rules is in the market basket analysis (MBA). Marketing research has been directed to the analysis of the co-incidence of multiple categories of products in different shopping, in order to plan marketing activities so that maximum profit is achieved. Retailers typically have to make decisions about which products put on sale, how and when.

According to Solnet et al. (2016) the basic idea underlying Market Basket Analysis is that consumers rarely make purchase decisions that are isolated. For example, when shopping in a supermarket, customers rarely buy one product; they are far more likely to purchase an entire basket of products, typically from different product categories. Using information about peoples market baskets allows data analysts to not only extract which products and product categories tend to be purchased together, but also to determine which of the products or product categories are drivers for purchasing certain products. This knowledge enables managers to develop interventions aimed at influencing purchasing behavior, including stimulating demand overall, promoting specific product categories, or offering promotions for driver products which are likely to increase overall spending per purchase.

According to Chen et al. (2005) the market basket analysis is a useful method of discovering customer-purchasing patterns by extracting associations or co-occurrences from stores transactional databases. Because the information obtained from the analysis can be used in forming marketing, sales, service, and operation strategies, it has drawn increased research interest. The research and discovery, for example, that supermarket customers are likely to purchase milk, bread, and cheese together, can help managers in designing store layout, web sites, product mix and bundling, and other marketing strategies.

According to Groth (2000), the MBA can be applied to: cross-selling analysis; layout definition; product catalogs design; leadership loss analysis; definition of price and product promotions; among others. These applications are based on the belief that sales of different product categories are correlated. For example, a promotion of beers could increase the sale of peanuts.

3 Data Modeling

Following the model proposed by Fayyad et al. (1996), a supermarket was selected and all its transactions were stored in a relational database where, through SQL

(Structured Query Language) was possible to perform advanced queries and then obtain the necessary information needed, for example: year, month, day of month, day of the week and the time the purchase was made and what products were obtained, what was the quantity purchased, the payment method, etc. A four months data window was selected (January, February, March and April). Table 1 shows the extracted information of each purchase made in this period.

Table 1 Extracted Information from Supermarket database

Information (*Y*)	Description (*X*)
Month	January, February, March or April
Month day	1, 2, ..., 31
Weekday	Sunday, Monday, Thursday, ..., Saturday
Day period	Morning, Afternoon, Evening
Day type	Normal, Holiday Eve, Holiday or Post-holiday
Value of Purchase	Total purchase price in Reais (BRL)
Shopping Basket Items	Example: apple, carrot, ...
Categories	Example: Fruit, Chocolates, ...

After the construction of the data file, a manipulation of the data was done in order to extract some basic information characterizing the purchases that are made at the supermarket and also to verify the quality of the data that have been selected and extracted. A table was created from the data file, where each line represents a purchase and each column the variables of this purchase (month, day, value, products, categories, etc.).

The first information retrieved from database was the volume of purchases made during supermarket first quarter. It was also noted that the purchasing volume changed according to the day time: morning, afternoon and evening. It has been found that there is an increased purchases in the morning, then afternoon and finally the period of night with a smaller volume.

By observing the total value of considered purchases made in these four months, it was revealed that approximately 96% of these purchases (245.042) are less than or equal to R$200.00 and are responsible for approximately 62% of total sales in the period. It is a supermarket characterized by small purchases. Approximately 90% of purchases are less or equal to R$100.00 and are responsible for approximately 43% of the total revenue.

Table 2 Example of table created from the data file

Month	Day	Weekday	Period	Type	Value	Product 1	Category 1
January	2	Friday	Morning	Pos-h	23,84	Coca Cola 2L	Soft Drink
January	2	Friday	Morning	Pos-h	16,94	Hot Dog Sausage	Bakery
January	2	Friday	Morning	Pos-h	169,2	Cocamar Soybean	Oils
January	2	Friday	Morning	Pos-h	118,53	Mant. Orange 2L	Soft Drink
January	2	Friday	Morning	Pos-h	67,21	Sucrilhos	Cereal
January	2	Friday	Morning	Pos-h	95,24	Floresta Coffee	Coffee
January	2	Friday	Morning	Pos-h	156,56	Onion 500g	Vegetable
January	2	Friday	Morning	Pos-h	163,98	Ades Peach	Juices
January	2	Friday	Morning	Pos-h	2,35	Red Eggs	Bakery
January	2	Friday	Morning	Pos-h	10,47	Fructis Cond.	Hair Conditioner
January	2	Friday	Morning	Pos-h	3,82	Bread	Bakery
January	2	Friday	Morning	Pos-h	1,79	Pineapple Sweet	Candy
January	2	Friday	Morning	Pos-h	76,76	Bauducco Toast	Toast
January	2	Friday	Morning	Pos-h	210,06	Knuckle	Beef

In the analyzed period, a total of 2,398,050 products was purchased with 13,114 different products in 184 different categories.

The Apriori algorithm was used to generate the association rules and find all frequent k-itemsets contained in a database. This algorithm generates a set of candidate k itemsets and then seek the database to determine if they are frequent, thereby identifying all frequent k-itemsets. The main feature of this algorithm is its downward closure. Through the apriori-gen function, the algorithm seek the database looking for the of frequent 1-itemsets, that is, those itemsets with only one item, and that satisfy the minimum support.

The next step is the discovery of 2-itemsets that satisfy the minimum support. Now, instead of the algorithm go through the entire database, it covers only the frequent 1-itemsets discovered in the previous step, as the support is always the same and then the 2-itemsets can only come from above. This procedure is based on the fact that an x-itemset has minimal support, then all subsets also have to. Likewise are generated 3-itemsets and so on. This property makes that is not necessary to go through the entire data set thus optimizing the generation task of frequent itemsets.

The association rule mining algorithms generates as a result rules like:

<beer → peanut; sup=20%; conf=40%>

This means that in 20% of all purchases made, beer and peanuts were purchased together. Considering now only purchases in which there was beer, in 40% of these also occurred peanuts.

The number of rules generated depends on the amount of purchase, quantity and attributes considered, specified as support and minimum confidence. In most cases this number becomes impracticable to observe all the generated rules.

For post-processing in order to transform the information into knowledge, the lift was used as evaluation measure. This measure also known as interest, is one of the most used to evaluate dependencies between itemsets (Silverstein et al., 1998). Given an association rule A → B, this measure indicates the more frequent becomes B when A occurs. The lift valor is calculated by equation 1:

$$lift(A \Rightarrow B) = \frac{conf(A \Rightarrow B)}{sup(B)} \quad (1)$$

If lift(A → B) = 1, then A and B are independent. If lift(A → B > 1), then A and B are positively dependent. If lift(A → B) < 1, A and B are negatively dependent. This measure varies between 0 to infinite, and has quite simple interpretation: the higher the lift, the more interesting the rule is, because A lifted B at a higher rate. For example, it means that B has 5 times more likely to occur when A occurs.

The lift ≥ 2 parameter was set to include only the interesting rules, i.e. rules were only considered in which B had at least twice more chance to happen when A occurs. This has reduced the number of interesting rules but it was necessary to also observe the support and confidence of the rule to choose an interesting set of rules.

The final rule set was obtained through trimming, specifying a minimum lift equal to 2 and also by observation of support and confidence values of the rules. Knowledge was generated from rules showing some behavior unknown by experts and other with obvious behavior or already known, but still rather interesting to measure.

4 Results and Discussion

The aim of this study was the market basket analysis of purchases by mining association rules on transactional data from a supermarket in order to provide greater insight into the buying behavior of their customers and discuss the applicability of the technique. The existence of a structured database, noise free and organized proved crucial to the analysis. A database with these characteristics facilitates the pre-processing of the information and makes it possible to generate quality results that accurately represent customer behavior.

This research applied the association rule mining technique in modeling of data from a typical supermarket in order to generate knowledge. A large number of patterns that could be used to decision aid have been found. The association mining rules was most helpful when a specific problem was proposed reducing the number of rules. This provided a focus to the analysis, reducing the search field of information and reducing the number of generated patterns.

A problem encountered in this type of study is that the reporting (and the execution method) and the decision-making are usually not made by the same person. Usually, the responsible for generating the report does not know what information

is crucial to the decision maker and does not know how to extract really useful information to the decision-making process; spending time and money to generate some useless reports. The association rules technique also has the advantage of being easy to interpret. It only takes a few basic statistical fundamentals to understand what can be extracted as a result of an application. The most difficult is to know what information is interesting for making a decision.

In short, the association rule mining technique provides an accurate summary of how the items are related. For future work a great inclusion to be explored is the use of loyalty cards where the customer can be identified in terms of age, sex, marital status, income, address, among others characteristics; a multivariate Bayesian forecast could also provide interesting insights about the stock and sales.

Acknowledgement

The authors thanks CAPES, CNPq and Fapemig for the support in this research.

5 References

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases. VLDB '94. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 487–499.

Chen, Y.-L., Tang, K., Shen, R.-J., Hu, Y.-H., Aug. 2005. Market basket anal- ysis in a multiple store environment. Decis. Support Syst. 40 (2), 339–354.

Fayyad, U., Piatetsky-shapiro, G., Smyth, P., 1996. From data mining to knowl- edge discovery in databases. AI Magazine 17, 37–54.

Grewal, D., Levy, M., 2007. Retailing research: Past, present, and future. Journal of Retailing 83 (4), 447 – 464.

Groth, R., 2000. Data Mining: Building Competitive Advantage. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Hipp, J., Gu¨ntzer, U., Nakhaeizadeh, G., 2002. Data mining of association rules and the process of knowledge discovery in databases. In: Industrial Conference on Data Mining: Advances in Data Mining, Applications in E-Commerce, Medicine, and Knowledge Management. Springer-Verlag, London, UK, UK, pp. 15–36.

Piatetsky-Shapiro, G., 2007. Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from "university" to "business" and "analytics". Data Mining and Knowledge Discovery 15 (1), 99–105.

Silverstein, C., Brin, S., Motwani, R., 1998. Beyond Market Baskets: General- izing Association Rules to Dependence Rules. Data Mining and Knowledge Discovery 2 (1), 39–68.

Solnet, D., Boztug, Y., Dolnicar, S., 2016. An untapped gold mine? exploring the potential of market basket analysis to grow hotel revenue. International Journal of Hospitality Management 56, 119 – 125.