Expert Systems With Applications 97 (2018) 146–162

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Market basket analysis: Complementing association rules with minimum spanning trees

Mauricio A. Valle[a,*], Gonzalo A. Ruz[b], Rodrigo Morrás[c]

[a] Facultad de Economía y Negocios, Universidad Finis Terrae, Santiago, Chile
[b] Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibáñez, Av. Diagonal las Torres 2640, Peñalolén, Santiago, Chile
[c] Escuela de Negocios, Universidad Adolfo Ibáñez, Av. Diagonal las Torres 2640, Peñalolén, Santiago, Chile

ARTICLE INFO

Article history:
Received 20 July 2017
Revised 24 November 2017
Accepted 14 December 2017
Available online 15 December 2017

Keywords:
Market basket analysis
Minimum spanning tree
Network of products
Association rules

ABSTRACT

This study proposes a methodology for market basket analysis based on minimum spanning trees, which complements the search for significant association rules among the vast set of rules that usually characterize such an analysis. Thanks to the hierarchical tree structure of the subdominant ultrametric distances of the MST, the association network allows us to find strong interdependencies between products in the same category, and to find products that serve as accesses or bridges to a set of other products with a high correlation among themselves. One relevant aspect of this graph-based methodology is the ease with which pairs and groups of products susceptible to carrying out marketing actions can be identified. The application of our methodology to a real transactional database succeeded in: 1. revealing product interdependencies with the greatest strengths, 2. revealing products of high importance with access to another product set, 3. determining high quality association rules, and 4. detect clusters and taxonomic relations among supermarket subcategories. This is highly beneficial for a retail manager or for a retail analyst who must propose different promotion and offer activities in order to maximize the sales volume and increase the effectiveness of promotion campaigns.

1. Introduction

Certain aspects of the administration of a supermarket chain can be critical to the success of sales and the chain's sustainability. For example, store managers must handle the inventory carefully, identifying which items sell the best and which tend to remain longer on the shelf. This is important for negotiating with suppliers.

In order to increase sales, a well-considered location for items can help improve the purchase volume. Some items are displayed at the customers' eye level, or high-turnover items (for example, bread and milk) are placed at the back of the store. What set of products would be good candidates for this type of strategy? What products or set of items should be part of discounts in exchange for a privileged space in the store? These are typical questions for a retail manager to keep the business competitive.

One way to achieve this is to take advantage of Expert Systems that can process data, extract knowledge and patterns from the enormous amount of accumulated data from customer transactions, and then, represent it understandably for a human decision-maker in a user-friendly way. An advantage of this approach is that it is easy to detect customer purchase behavior patterns, which would otherwise be impossible. Thus, specific and targeted promotion strategies based on empirical data are within the manager's reach.

The market basket analysis (MBA) deals precisely with finding consumer purchase patterns from transactional databases (Linoff & Berry, 2011). MBA focuses on discovering buying patterns along thousands or millions of transactions. In this task, Association Rules (ARs) have played an important role to find frequent itemsets and relationships of purchases between different products in which explicit declarations of the type are established: *if item X is bought, then item Y is also bought* (Agrawal, 1994).

One of the disadvantages of association rules is that even with databases of hundreds of transactions, the algorithm produces a very high number of rules which makes it impossible to analyze them all (Kotsiantis & Kanellopoulos, 2006). The traditional way to overcome this problem is by imposing a restriction on the basic measures of the ARs, for example, leaving aside all those rules with support and confidence below a predefined threshold. Even with this solution, however, the number of rules is still too large. Another issue with ARs is that they usually produce rules that are spurious (García, Romero, Ventura, & Calders, 2007), so

* Corresponding author.
E-mail addresses: mvalle@uft.cl (M.A. Valle), gonzalo.ruz@uai.cl (G.A. Ruz), rmorras@uai.cl (R. Morrás).

https://doi.org/10.1016/j.eswa.2017.12.028
0957-4174/

it is difficult to identify which rules are practical and useful for decision-making.

An alternative to association rules is transactional data modeling in the form of networks. The nodes that comprise the network are items or product categories, while the edges that unite the nodes represent the simultaneous occurrence of two products. These kind of networks are called network of products (Raeder & Chawla, 2009). This way network-based techniques can be applied, acquiring valuable data on customer purchase behavior, which is very difficult with association rules. The network-based approach isolates the influence between products, which can detect the interrelation of those that are strongest through community detection methods widely recognized in the literature. For example, Raeder and Chawla (2011) developed a product community measurement based on the confidence of the relationship indicated by the network edge. This makes it possible to find groups of items that are highly likely to be bought at the same time.

Graph mining techniques have the advantage of being able to process large volumes of information and displaying the results easily and intuitively using intensity thresholds and filters on the edges of the product network (Ríos & Videla-Cavieres, 2014; Videla-Cavieres & Ríos, 2014). Thus, product associations can be more easily interpreted. An interesting extension of the market basket network is the co-purchased product network. In this network, items are linked when they are bought by the same customer (Kim, Kim, & Chen, 2012). The analysis examines customer purchase behavior, promising to establish personalized services and purchase recommendations aimed at certain types of customers more efficiently.

On the other hand, graphical techniques exist that help the visual exploration of ARs. Some authors have proposed ARs visualization techniques to help deal with the large number of rules that the algorithm generates. For example, *grouped representation* has been proposed (Hahsler & Karpienko, 2017) in which a significant number of rules can be viewed in a matrix form or in a graph-based visualization. Another approach is to use *parallel coordinates approach* (Bruzzese & Buono, 2004) in which multidimensional data sets are transformed in such a way that it is possible to observe the ARs in a 2D graph. Representations in $n \times n$ grids, in which the LHS and RHS of the rules labels the rows and columns, are useful for observing and comparing association rules, especially when the graph allows interaction with the user (Sekhavat & Hoeber, 2013). Another alternative arises when deploying clustered association rules in a 2D dimentional graph, observing rules characterized by certain levels of measures of quality (Couturier, Hamrouni, Yahia, & Nguifo, 2007; Kim, 2017).

Although useful, these graphical representations work directly with the rules already generated, so that they can be only used as means to compare other rules according to a quality measure such as confidence and lift. We focus in using minimum spanning trees (MST), which allows us to observe how the products or items are related to each other, and not as part of a rule. This advantage allows the analyst to be active in the search for interesting rules for the retail manager, observing at first-hand stronger relations between products, and from there, to look for rules found by the Apriori algorithm that involve those products of interest. In other approaches, the analyst confines himself to finding rules with a better quality attribute.

The MST allows us to reduce the complexity of the network of products by interconnecting the nodes with the highest correlation and discarding low co-occurrence or random interconnections. The MST has been an important part of data mining and machine learning applications. For this reason, researchers have developed new and faster algorithms and variants of the Prim and Kruskal algorithm to find MST for large amounts of available information. For example, by using Artificial Neural Networks (Ferilli, Sacco, Teti, & Buscema, 2016; Wang, Wang, Ma, & Wilkes, 2015; Zhong, Malinen, Miao, & Fränti, 2015), others using KNN (Zhong et al., 2015), others using fuzzy level weight coefficients (Gao, Zhang, Lu, Wu, & Du, 2017), and also to learn extended version of the tree-augmented naive Bayes classifiers (de Campos, Corani, Scanagatta, Cuccu, & Zaffalon, 2016).

The input information to the proposed method is the correlation matrix between the set of all product vectors that denote the presence or non-presence of them in consumer market baskets. These correlations are directly related to the lift, which measures the quality of association rules, indicating the level of association between antecedent and the consequence of the rule. The output consists of an undirected graph with at most N nodes and $N-1$ edges connecting the products with high level of interdependencies (or minimum distance path). This is the MST that can be seen itself as the set of rules of maximum quality, and from which more complex rules can be searched that involve more number of items in the rules.

The main goal of this paper is to propose a methodology to analyze the structure and behavior of consumer market baskets from the discovery of the network topology. On the basis of the methodology, we develop two specific elements that guide the effective search for association rules: The first one is the detection of strong links between products, which are represented in the MST as smaller distances relative to the rest of the edges in the tree. The second one, is the detection of key products by a simple measure of local importance that takes into account the degree and the connection distances of the node to other adjacent ones. As a result, certain areas of the spanning tree (with low distance between nodes, and nodes with access to other products) provide to the analyst, potential action points for effective promotional activities.

The main novelty in this study is the simplicity with which it is possible to visualize or report the product sets with a high level of complementarity or co-occurrence from large volumes of data contained in the hundreds of thousands of customer transactions in a certain period. This approach responds to the need to offer a robust visualization for the human analyst, which is in line with the work by Techapichetvanich and Datta (2004). To be more specific, our proposal enables:

1. The discovery of the strongest interdependencies between products of the same and different categories. For example, given product A, finding (if there is one) a product set S that has a strong dependency on A, i.e., both appear in the same market basket. This is equivalent to focusing the search for *high quality* association rules.
2. The identification of key products that act as the nexus to other different product groups of the same or a different category, and the recognition of zones of influence in the MST. This makes it possible to use the methodology as a pruning tool for narrowing the scope of ARs.
3. The use of the network of products from which it is possible to reduce the complexity of the analysis by building the associated minimum spanning tree. Consequently, the resulting MST gives us the hierarchical structure of the products under analysis.
4. The analysis of taxonomic groups revealed by the ultrametric distances of the MST, which allows us to study the coherence between the product categories used by the supermarket, and the underlying product groups formed from the purchasing behaviour.

Our model is based on the level of correlation of items present or not present in the customers' market baskets. Therefore, the focus is on discovering the interdependencies between items or product subcategories and finding those that serve as bridges to different product groups. This view of the issue through correla-