International Conference on Advances in Business and Law
Proceedings Series, ICABL, 2019, Vol. 3, ISSN: 2523-6547
Available at https://publications.ud.ac.ae/index.php/ICABML-CP

3rd International Conference on Advances in Business and Law (ICABL-2019)
23-24 November 2019, Dubai, UAE

Hierarchical clustering of products using market-basket data

Ondřej Sokol [a]*

[a]*University of Economics, Prague, Czech Republic, ondrej.sokol@vse.cz*
**Corresponding author.*

ABSTRACT

The goal of this paper is to present a new method of clustering products based only on the market-basket data from the retail store. The presented approach uses a special way of computing the dissimilarity matrix on which Ward's hierarchical clustering method is used. The similarity matrix stems from the co-occurrence of products in same basket as a utility data. As a similar are denoted products which have similar co-occurring products and simultaneously are not often present in the same basket. Hence, the method does not require the identification of the customer, neither the data from fixed time frame, which is an advantage over commonly used methods. The method is reasonably fast even over huge dataset of tens of millions rows. The results are promising and easy to interpret.

Keywords: *product clustering, market basket data, hierarchical clustering, retail.*

JEL codes**:** *C38, L81*

1. INTRODUCTION

Data mining becomes more important as the companies are able to gather more data. Every retail company generates a huge amount of data every day and as the data amount raises, the data mining in sense of deriving information purely from the data becomes more important. A correct understanding of the data allows improving the business decision-making process. In this paper, we propose a new method to cluster products using only market basket data.

We focus on the product categorization which is area very important mainly in marketing, e.g. new product development (Gruca, 2003), analysis of cross-category sales promotions (Leeflang, 2008) or optimizing placement of retail products on shelves (Borin, 1994). Another utilization is the replacement of the product which runs out. Sold-out products are usually replaced by other "similar" ones.

Products are usually categorized by their purpose and package properties such as brand, package size and price. Another approach was presented by (Srivastava, 1981) who used hierarchical clustering based on substitution-in-use criteria and (Zhang, 2007) who promoted fuzzy clustering.

Market basket data are usually used for analysis of cross-category dependence for a priori given categories (Russell, 2000). In (Holý, 2017) we presented a method to cluster products based on their co-incidence in the basket using a genetic algorithm. In this paper, we present a method to the hierarchical categorization of products based only on the market basket data. The resulting clusters group products with similar "other" products in their baskets. The main idea is that two similar products have similar affinity indices in regards to the all other products. Aggregation of such indices allows to perform clustering using common techniques such as Ward's hierarchical clustering. In an application to a Czech drugstore chain, we describe the resulting hierarchical clustering on the category of drugstore creams.

http://dx.doi.org/10.30585/icabl-cp.v3i1.488

http://dx.doi.org/10.30585/icabl-cp.v3i1.488

The paper is organized as follows. In Section 2 we define the goal and basic comparison to common approach. In Methodology section, we describe our setup and the method of computing similarity (and dissimilarity) between products. Then in section Data and Findings we depict the real dataset for our application and discuss the resulting clustering. The paper concludes in the section Conclusion.

2.GOAL

The substitute measure SM_{ij} between products P_i and P_j (i, j are indices of products) are usually computed using data which are linked to customers (e. g. customers must have loyalty card or other identification) such as

$$SM_{ij} = \frac{cust([P_i, P_j])}{cust(P_i)cust(P_j)}$$

where $cust(P_i)$ is the number of customers who bought product P_i during some given time frame and $cust([P_i, P_j])$ is the number of customers who bought both products P_i and P_j during the given time frame. The disadvantage of the method is the necessity of the identification of the customer and necessity of having data from longer time frame (to ensure more visits of the store of customers). This is not a problem in some areas, e. g. e-shop; however, it is rather uncommon to have reliable identification of customer in retail industry.

We propose a method for estimation of substituting products using every receipt from retail dataset – we are not restricted only to the receipts with the customers. Even more, we are not restricted to given time frame as the method does not depend on time of the purchase. Instead we use every single receipt to compute the similarity using their co-occurrence. The methodology is described in Section 3.

3. METHODOLOGY

We use common hierarchical clustering approach for points in R^n. Basic review of clustering method can be found in (Jain, 1999). What is novel is the way of computing similarity between points – in our case each point $x_i \in R^n$ for $i = 1, \dots, n$ represents a product. The overview of approaches to compute similarity between objects can be found in (Cha, 2007).

We focus on Ward's method for hierarchical clustering and similarity computed using the information about co-occurrence of products in joint baskets. In the next subsections our approach is described along with basic comparison to the commonly used methods.

3.1 Ward's method of hierarchical clustering

Ward's method (Ward, 1963) is well-known criterion used in hierarchical cluster analysis. The goal of the Ward's method is to find hierarchical clustering which minimize the total within-cluster variance – this also means maximizing the between-cluster variance.

Assume x_i are points in R^n for $i = 1, \dots, n$. Then within-cluster variance $W(C^k)$ for clustering C^k of k clusters is computed for Euclidean norm as

$$W(C^k) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left| x_i - \overline{x_j} \right|^2$$

where C_j is the set of points which are assigned to j-th cluster and $\overline{x_j}$ is the mean of cluster C_j.

The points x_i are defined by the dissimilarity matrix. There are many ways to compute the dissimilarity matrix, see (Cha, 2007). In our case, we compute similarity using formula in Section 3.2.

At the start of the algorithm, all points x_i are in exclusive clusters, therefore all clusters are singleton. The algorithm is iterative and in each the step a pair of clusters is merged while the objective function is minimized in regards to the lowering number of clusters.

http://dx.doi.org/10.30585/icabl-cp.v3i1.488

3.2 Computing dissimilarity over market-basket data

Consider dataset of m receipts where m is a very large number (usually more than tens of thousands). Each receipt R is a set of products P.

Define a_i for $i = 1, \dots, n$ as the number of occurences product P_i appeared on (distinct) receipts. Similarly, define a_{ij} as the number of occurrences that both product P_i and P_j appeared on the same receipt.

It is needed to consider that products are not sold at the same rate, hence we need to normalize the data. Define

$$A_{ij} = \frac{a_{ij}}{a_i}$$

This is a known formula for computing product affinity, see for example (Lallich, 2007). However, this is only first step. In our approach, we are not interested in affinity solely between two products, but their affinity computed as a sum of affinity indices over any other products.

We can then count similarity s_{ij} between two products P_i and P_j using the similarity of their common joint products on the receipt. This can be written as a sum of normalized co-occurrences $A_{ik}A_{jk}$ over all other products k.

$$s_{ij} = \sum_{\substack{k=1 \\ k \neq i,j}}^{n} A_{ik}A_{jk}$$

The similarity is an index in the range of 0 and n with higher value means the similarity of products is greater. In our real drugstore retail dataset, the products with high similarity had s_{ij} barely over 1. Those values who exceed 1 are assigned to 1; however, this was a case for a few values s_{ij}. This is an important step as the results would not be interpretable without mentioned normalization. We use a property of our dataset which allows us to do this normalization in which we lose very little information given how rare the condition for it occurs. Another approach would be to normalize the whole matrix of s_{ij} to ensure the highest value is equal to 1.

Having computed similarity index for all combination of products i and j we are able to use common hierarchical clustering methods. In our case, we used Ward's method implementation (ward.D2) in R hclust package. The method works with dissimilarity index instead of similarity index, therefore we use transformation $d_{ij} = 1 - s_{ij}$ as the method's input. The output of the method is hierarchical clustering of all products, e. g. clustering for each level from 1 to n. The goal is to find clusters of products with the similar perception by the customers.

Another approach is to introduce the penalization for occurrences of P_i and P_j on the same basket. The idea stems from the economic theory of substitutes. If P_i and P_j are substitutes, then they should not occur on the same receipt frequently. However, the co-occurrence with the different products should be similar. For this reason, we can compute the similarity of two products using the following formula which involves penalization for the occurrences in the same basket.

$$s'_{ij} = \sum_{\substack{k=1 \\ k \neq i,j}}^{n} \left(A_{ik}A_{jk}\right) - cA_{ij}A_{ji}$$

where c is a constant which defines the weight of the penalisation. The natural choice would be $c = 1$ which we use in the following experiments. Similarly to the previous method we use Ward's method implementation (ward.D2) in R hclust package with $d'_{ij} = 1 - s'_{ij}$ transformation to dissimilarity measurement.

3. DATA

We used a dataset from the retail drugstore chain. We focused on skin care products which is a group of categories like day, night and specialized skin creams, body milks, lip balm and cleansing products. The group consists of 533 products after data cleansing during which we omitted products with very low sales (less than 100 in a year).

http://dx.doi.org/10.30585/icabl-cp.v3i1.488

The dataset of over 10 million receipts was used. In the first phase, we computed a_i of skin products. Then we computed a_{ij} in which are included other products as well in the sense that i is restricted to only skin product while j is unrestricted. This phase was done using SQL data warehouse and the computation took approximately 5 minutes.

The second phase was done in R software using which we run both formulas for hierarchical clustering, with and without penalization for the occurrences in the same basket. The mentioned method is very fast in both cases and for $n \sim 500$ the results are given in a few seconds.

4. ANALYSIS AND RESULTS

Some clusters are formed along brands, some along category. This means that in some cases the brand is more important than category. For example, customers prefer to buy the slightly different colour of hair dyes than change their favourite brand. This finding brings new information about product position which may be hard to estimate by experts. But in our case it is also true that the resulting clustering is easy to understand.

What is also important is that both presented method (with or without penalisation term) behave differently. First method clusters products bought by similar customers, while the other is using additional information of possible substitution effect. For the illustration of the mentioned differences, we show tree cut of hierarchical clustering at level 70 on a case of L'Oréal Age products (at level ~30 all L'Oréal Age products were in the same cluster and we are interested in further subclustering). In Table 1 is resulting clustering of the method without penalisation of substitutes and in Table 2 is resulting clustering of the method with penalisation.

Table 1. Clustering of L'Oréal Age creams at tree cut level 70 without penalization

Product	Cluster
L'Oréal Age DUO 65+ 50ml	A1
L'Oréal Age DUO 55+ 50ml	A1
L'Oréal Age Eye Cream 55+ 15ml	B1
L'Oréal Age Day Cream 35+ 50ml	C1
L'Oréal Age Night Cream 35+ 50ml	C1
L'Oréal Age Day Cream 45+ 50ml	C1
L'Oréal Age Night Cream 45+ 50ml	C1
L'Oréal Age Day Cream 55+ 50ml	C1
L'Oréal Age Night Cream 55+ 50ml	C1
L'Oréal Age Day Cream 65+ 50ml	C1
L'Oréal Age Night Cream 65+ 50ml	C1

Products are clustered into 3 groups. Cluster A consists of two DUO products which are packages of day and night creams. Eye cream product is exclusively in cluster B and all day and night creams are in cluster C regardless of the recommended age.

Clustering is similar to the method without penalization; however, the products which were assigned to cluster C1 are divided into two clusters C2 and D2. Day and night creams are commonly bought together, therefore they are not in the same cluster with respect to the recommended age. It is interesting that C2 consists of 3 night creams and 1 day cream and D2 has the exact opposite distribution. The algorithm can reliably estimate complements (in this case day and night cream with same recommended age) but clustering substitutes is not perfect (yet).

http://dx.doi.org/10.30585/icabl-cp.v3i1.488

Table 2. Clustering of L'Oréal Age creams at tree cut level 70 with penalization

Product	Cluster
L'Oréal Age DUO 65+ 50ml	A2
L'Oréal Age DUO 55+ 50ml	A2
L'Oréal Age Eye Cream 55+ 15ml	B2
L'Oréal Age Day Cream 35+ 50ml	C2
L'Oréal Age Night Cream 45+ 50ml	C2
L'Oréal Age Night Cream 55+ 50ml	C2
L'Oréal Age Night Cream 65+ 50ml	C2
L'Oréal Age Night Cream 35+ 50ml	D2
L'Oréal Age Day Cream 45+ 50ml	D2
L'Oréal Age Day Cream 55+ 50ml	D2
L'Oréal Age Day Cream 65+ 50ml	D2

4. CONCLUSIONS

We presented a method for hierarchical clustering of products based on their substitution relationship. The only inputs of both methods are market basket data and the co-occurrence of the products in baskets while expert opinions or other qualitative information is not involved which is an advantage of the method. We cluster products which are bought by similar customers while adding the information about possible substitution effects. The resulting hierarchical clustering gives promising results which are easy to interpret.

FUNDING

The work on this paper was supported by IGS F4/78/2018, University of Economics, Prague.

REFERENCES

Borin, N., Farris, P. W., & Freeland, J. R. (1994). A Model for Determining Retail Product Category Assortment and Shelf Space Allocation. Decision Sciences, 25(3), 359–384. https://doi.org/10.1111/j.1540-5915.1994.tb00809.x

Cha, S. H. (2007). Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. International Journal of Mathematical models and Methods in Applied Sciences, 1(4), 300-307.

Gruca, T. S., & Klemz, B. R. (2003). Optimal new product positioning: A genetic algorithm approach. European Journal of Operational Research, 146(3), 621–633. https://doi.org/10.1016/S0377-2217(02)00349-1

Holý, V., Sokol, O., & Černý, M. (2017). Clustering Retail Products Based on Customer Behaviour. Applied Soft Computing. https://doi.org/10.1016/j.asoc.2017.02.004

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323.

Lallich, S., Teytaud, O., & Prudhomme, E. (2007). Association rule interestingness: Measure and statistical validation. In Quality measures in data mining (pp. 251-275). Springer, Berlin, Heidelberg.

Leeflang, P. S. H., Parreño Selva, J., Van Dijk, A., & Wittink, D. R. (2008). Decomposing the sales promotion bump accounting for cross-category effects. International Journal of Research in Marketing, 25(3), 201–214. https://doi.org/10.1016/j.ijresmar.2008.03.003

Russell, G. J., & Petersen, A. (2000). Analysis of cross category dependence in market basket selection. Journal of Retailing, 76(3), 367–392. https://doi.org/10.1016/S0022-4359(00)00030-0

Srivastava, R. K., Leone, R. P., & Shocker, A. D. (1981). Market Structure Analysis: Hierarchical Clustering of Products Based on Substitution-in-Use. Journal of Marketing, 45(3), 38. https://doi.org/10.2307/1251540

Ward, J. H., Jr. (1963). Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association, 58, 236–244.

http://dx.doi.org/10.30585/icabl-cp.v3i1.488

Zhang, Y., (Roger) Jiao, J., & Ma, Y. (2007). Market segmentation for product family positioning based on fuzzy clustering. Journal of Engineering Design, 18(3), 227–241. https://doi.org/10.1080/09544820600752781