Time Variations of Association Rules in Market Basket Analysis

Vasileios Papavasileiou and Athanasios Tsadiras

Department of Economics, Aristotle University of Thessaloniki
GR-54124 Thessaloniki, Greece
{vapapava,tsadiras}@econ.auth.gr

Abstract. This article introduces the concept of the variability of association rules of products through the estimate of a new indicator called overall variability of association rules (OCVR). The proposed indicator applied to super market chain products, tries to highlight product market baskets, with great variability in consumer behavior. Parameter of the variability of association rules in connection with changes in the purchasing habit during the course of time, can contribute further to the efficient market basket analysis and appropriate marketing strategies to promote sales. These strategies may include changing the location of the products on the shelf, the redefinition of the discount or even policy or even the successful of recommendation systems.

Keywords: Market Basket Analysis, Association Rules, Data Mining, Marketing Strategies, Recommendation Systems.

1 Introduction

1.1 Market Basket Analysis

The consumer behavior data collection, for a super market chain, via an appropriate strategy, may lead the company to significant economies of scale. For the same reason, the extension of the use of loyalty cards, constitutes a primary goal for the administration of a super market chain. In particular, the market basket analysis [1] can bring out the combinations of products that are susceptible of marketing strategies and may lead to better financial results. With market basket analysis [2], the administration of a super market can understand the behavior and purchasing habits of customers, through combinations of product market [3], which is repeated in large or small degree, as a habit. These combinations are called products association rules [4,5,6,7] and are the result of market basket analysis procedure.

The simplest form of an association rule, shows two products and has the form X > Y, which means that the purchase of product X leads to the purchase of product Y. For rules evaluation there are used three main indicators [8], which is the degree of confidence, the degree of support and the degree of lift. The degree of confidence expresses the possibility of realization of rule X > Y in the set of transactions involving the purchase of the product X. Respectively, the degree of support expresses the possibility of realization of rule X > Y in the set of all transactions.

L. Iliadis et al. (Eds.): EANN/AIAI 2011, Part II, IFIP AICT 364, pp. 36–44, 2011.

Finally, lift Indicates how much better the rule is at predicting the "result" or "consequent" as compared to having no rule at all, or how much better the rule does rather than just guessing.

1.2 Dataset

The data of this survey are collected from a Greek known super market chain and are related to annual customer transactions [9,10] for the period from 01/09/2008 to 31/08/2009. Data are sourced from the information system of the company and come exclusively from the customer card transactions through loyalty cards. Every transaction is a record of the purchase of specific products carried out by the customer who makes use of the card at the time of purchase.

As far as data distribution is concerned there has been observed the long tail effect, which belongs to the broader category of low power distributions that appear quite often in nature. The long tail effect has assumed its present connotation and dynamics by Chris Anderson [11]. Figure 1 below shows the curve of long tail effect.

This term often refers to data products purchase in supermarkets describing their distribution as a long tail in which a small number of products is purchased more frequently whereas a large one is purchased less frequently. This phenomeno creates data sparsity problem and worsens even more their elaboration. For this survey from the total number of transactions there were selected, those that included at least the purchase of 8 products. That is how, to some extent, there was tried to resolve the data sparsity problem and prevent the removal of any market basket product, which may present a risk of information loss.

During the stage of preprocessing [12,13,14], the products were grouped into categories and subcategories, depending on their utility and the classification in supermarket shelves. The number of codes of the products sold by the chain is approximately 95.000 and with the categorization applied there was diminished in 247 product subcategories.

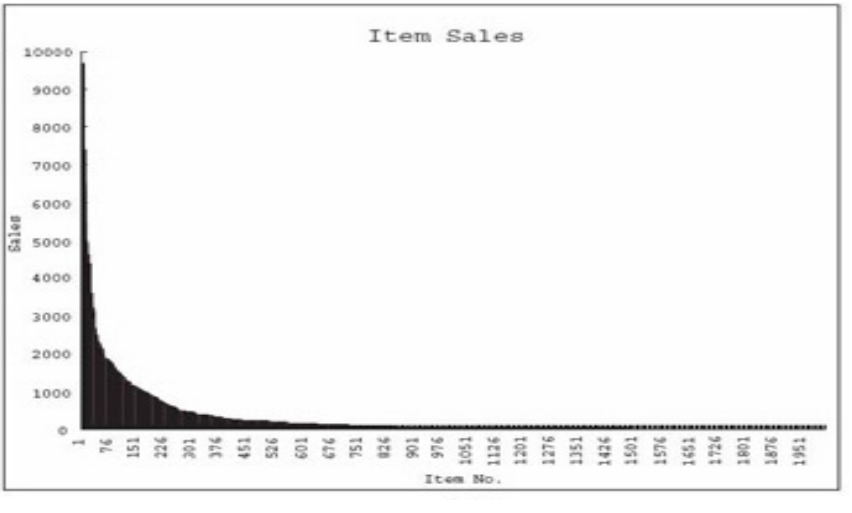

Fig. 1. Curve of Long Tail Effect

Table 1. Number of Transactions per Year and per Month

Time	Year	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6
Number	183.827	15.603	16.027	14.621	18.141	15.323	15.982
		Month 7	Month 8	Month 9	Month 10	Month11	Month 12
		16.189	15.642	15.907	14.624	13.189	12.579

Occasionally, there have been proposed various algorithms for efficient market basket analysis. The most widely established, is the a priori algorithm, delivered for the first time by Rakesh Agrawal and Ramakrishnan Srikant [15]. For the extraction of data there was used the free software Tanagra [16]. This specific data mining tool is specialized in market basket analysis and is applied in the a priori algorithm. For compatibility with the program, the sets of data were formatted in binary tables [17], where each row is a separate transaction and each column represents one of the 247 sub-categories of products.

Each product of a transaction corresponds to 1, when purchased and to 0 when it isn't. Hence, there were formed 12 binary transaction tables respectively to the 12 months of analysis and 1 binary transaction table for the annual purchase data. The market basket analysis was applied to each of the 12 subsets separately, in a subcategory level of products and the results were studied in a association with time. In table 1 there is the number of transactions elaborated, both for the whole year and the 12 sub-totals. With the indication Month 1 we refer to the total number of all purchasing data for September 2008, while for Month 12, it is implied the total number of purchasing data for August 2009 respectively.

2 Time Variations of Association Rules

2.1 Discoveries of Association Rules from Dataset

For all of the data and for each of the subsets there has been activated the procedure of market basket analysis and recorded the results of the three indicators, confidence, support and lift. For every application of a market basket analysis process, the minimum level of confidence, support and lift established in the program, is 0.25, 0.01 and 1.1 respectively. The results of the number of rules that came out, are registered in table 2 and correspond to the level of 247 product subcategories.

Table 2. Number of Annual and Monthly Association Data Rules

Time	Year	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6
Number	1.803	2.262	2.221	2.537	2.429	2.440	1.929
		Month 7	Month 8	Month 9	Month 10	Month11	Month 12
		1.619	1.875	2.149	1.696	1.651	2.215

2.2 Time Development of Association Rules

The 1803 rules, at the level of 247 subcategories of products, resulting from the extraction process of overall annual data were used as a base for additional rules analysis at monthly level. For each of the 1803 rules, there have been recorded the values for the three indicators, lift, support and confidence during the 12 months of the year. Thus was created the table 3, with the monthly development of association rules measurement indicators. In table 3, L indicates lift, S is for the value of support and C denotes the value of confidence.

In an attempt to show the price development of the degree of confidence in months 1, 2 and 12 it comes out the graph of time change of the degree of confidence of association rules.

Rules 5, 6 and 7 present a significant slope owing to the strong variability of the degree of customer confidence in September 2008, October 2008 and August 2009. It is also noted that when the confidence of customers is increased for the market basket of rule 5, the same period the confidence of customers for the market basket of rule 6 is reduced. The same but to a lesser extent seems to be the case with rules 5 and 7. In such cases the market baskets between them seem to work as substitute.

Table 3. Monthly Development of Association Rules Measurement Indicators

		Month 1 (Sept. 2008)			*Month 2 (Oct. 2008)*		
A/A	*Rules*	L	S %	C %	L	S %	C %
1	Yoghurt – Bread > Fruits	1.83	1.27	86.84	1.72	1.49	81.02
2	Toast- Cereals> Fresh Milk	1.20	2.26	71.89	1.23	2.52	73.81
3	Fruits – Sweet Biscuits > Vegetables	1.73	2.63	64.70	1.79	2.44	61.77
4	Fruits – Pasta > Fresh Milk - Vegetables	1.83	1.87	40.70	1.84	1.69	36.44
5	Fresh Milk – Beers > Cola	1.86	1.27	40.60	11.2	1.12	66.42
6	Cola – Salad (Not Packed, Packed) > Vegetables	12.8	1.02	79.50	1.14	1.41	68.28
7	Chicken Not Packed> Fruits	14.1	1.57	53.49	1.17	1.01	30.55
.	……..	…	…	…	…	…	….
.	……..	…	…	…	…	…	….
1803	Fresh Milk – Beef Not Packed> Bread	1.63	1.28	33.67	1.73	1.18	34.36
		….. .			*Month 12 (Aug. 2009)*		
		L	S%	C%	L	S%	C%
		…	…	…	1.85	1.26	85.95
		…	…	…	1.38	2.58	79.41
		…	…	….	1.56	2.70	62.62
		…	…	….	1.61	1.53	36.42
		…	…	…	1.69	1.71	37.79
		…	…	…	15.2	1.19	99.34
		…	…	….	1.16	1.26	32.78
		…	…	….	…	…	….
		…	…	….	…	…	….
		…	…	….	1.58	1.34	31.12

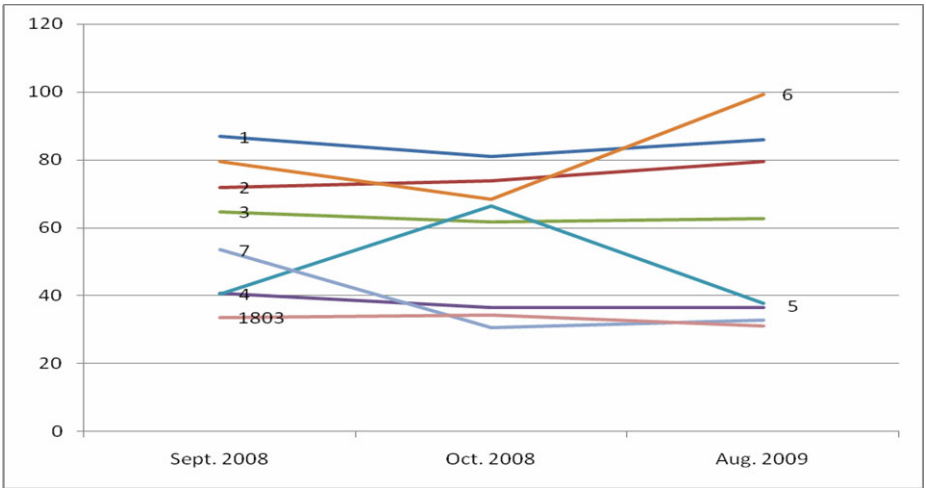

Fig. 2. Graph of the Time Change Confidence Degree of the Association Rules

As shown by the results in table 3, the values of association rules indicators lift, confidence, support, of super market chain products, present monthly fluctuations. For example, the degree of confidence in the rule 1 appears with percentage 86.84% in September and is reduced to 81.02% in the month of October and is increased to 85.95% the month of August. The option to purchase fruit after the selection of bread and yoghurt in the basket market seems to be less random in September compared to October and more random than that in August. Also on rules 5, 6 and 7, for the first three months under consideration, there are significant changes in levels of confidence in connection with the other rules. In particular, for rule 6 the changes are made to a high degree of confidence.

These monthly fluctuations in prices of lift and confidence are the subject of this research study. The research of rules association products should not be concluded solely on the assessment of the three indicators, without taking account of changes in these prices during the year. Through the assessment of variation in the degree of confidence and lift of a rule, one can draw useful conclusions about the marketing. The concept of variability has occupied the research of the export association rules geographical and temporal data within the study mainly climatic [18] and weather [19] phenomena. This article attempts to highlight the term of variability in critical factor in analyzing purchasing data of super markets.

3 Define of OCVR

For the evaluation of changes in the lift and confidence values of association rules, there has been calculated the statistical indicator of relative variability or standard deviation [20]. The indicator of the relative standard deviation, was chosen as the most effective for the comparison of the variability of the rules, as it is the most appropriate to compare the observations expressed in the same units but whose arithmetic average differ significantly. The variability index CV, is defined as the ratio of the standard deviation s by the mean value $\overline{X}$.

$$CV = \frac{s}{\overline{X}} \ . \tag{1}$$

For each of the 1803 association rules during the 12 months there has been calculated the index variability lift (CVL) and the index variability confidence (CVC).

In order to aggregate both CVL and CVC indexes and combine these indexes into a single index, we introduced a new indicator the one of the overall variability of association rules (OCVR), as a result of the average variability of lift and confidence.

$$OCVR = \frac{CVL \ + CVC}{2} \ . \tag{2}$$

4 Experimental Study

Table 4 lists the results of price variability index lift (CVL), the index variability confidence (CVC) and the index of overall variance (OCVR) specific association rules.

Table 4. Overall Variability Association Rules

A/A	*Rules*	*CVL %*	*CVC %*	*OCVR %*
1	Yoghurt – Bread> Fruits	4.19	4.46	4.33
2	Toast- Cereal > Fresh Milk	4.05	2.86	3.46
3	Fruits – Sweet Biscuits > Vegetables	5.40	4.27	4.83
4	Fruits – Pasta > Fresh Milk - Vegetables	5.11	5.66	5.39
.	……	…	…	…
.	……	…	…	…
.	……			
1803	Fresh Milk – Beef not Packed > Bread	11.10	24.80	17.95

We note that the indicator of overall variability of rule 1 is 4.33% and the corresponding value of confidence for the three months that were presented in table 3, ranges from 81.02% to 86.84%. The indicator of overall variability of rule 2 is 3.46 and the corresponding value of the confidence ranges from 71.89% to 79.41%. This could mean that the rule 1 which is realized by a greater degree of consumer confidence than rule 2, can be changed more easily compared to rule 2, which seems to be more compact and inflexible. The indicator of overall variability of rule 3 is 4.83 and the corresponding value of the confidence ranges from 61.77% to 64.70%. This could mean that the rule 3 which is realized with a lesser extent consumer confidence compared to rule 2, can be changed more easily than rule 2. Rule 3 compared to rule 2 could be perceived as a more sensitive and flexible.

Figure 3 below shows the curve of the overall variance ratio (OCVR) for all of 1803 association rules for products of super market, for the period from 01.09.2008 to 31.08.2009.

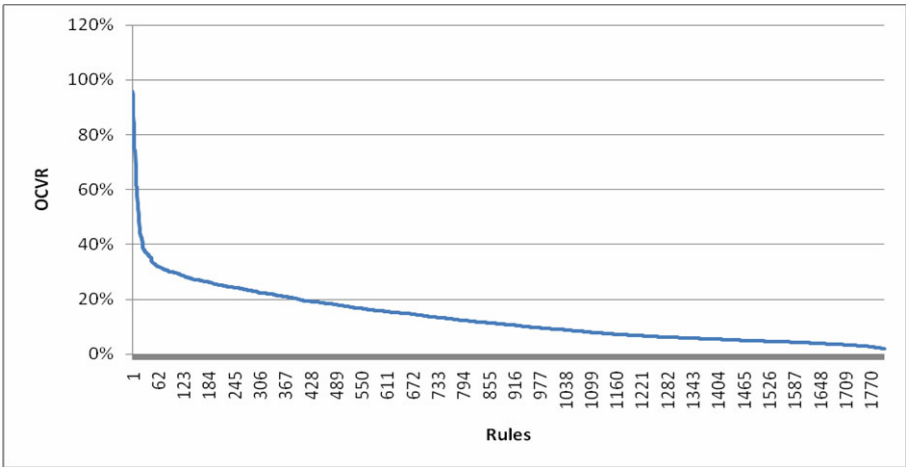

Fig. 3. Overall Variability Curve of Association Rules

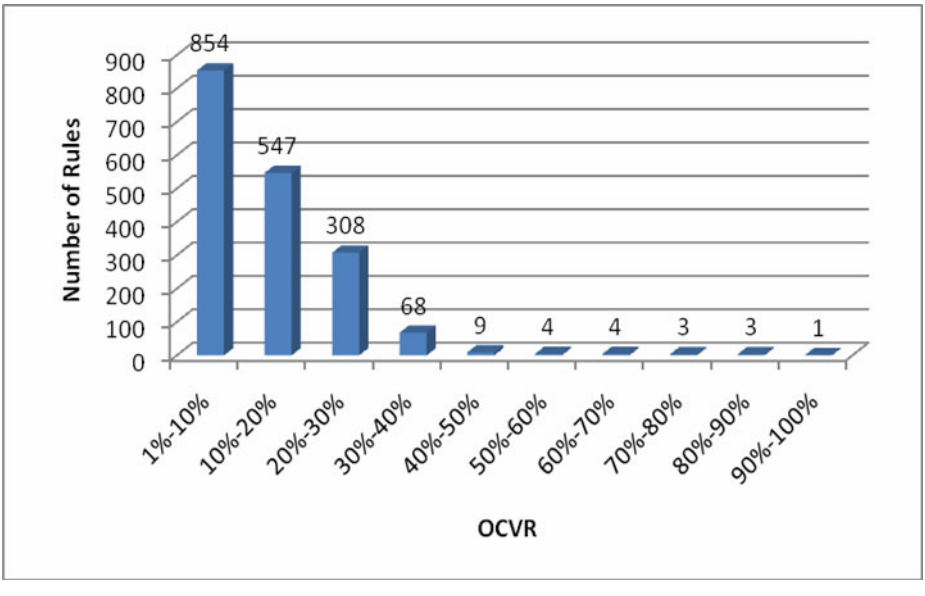

Fig. 4. Histogram of Overall Variability of Association Rules

We can observe that the form of the distribution of values of OCVR follows the long tail distribution. More specifically, a small number of rules presents a high overall variability. In Figure 4 is shown the histogram of the values of the overall variability. We observe that the majority of the overall variability index values range from 1% of space – 30% while there are very few rules between values 30%-100%.

5 Conclusions

The OCVR allows for the evolution and understanding of the association rules. There might be a particular interest in rules that receive high values of OCVR index. These rules seem to vary systematically and with intense degree, which means that the purchasing behavior of consumers in this particular market basket is not so stable and predefined. However, this does not mean that these rules do not have value for further analysis. On the contrary, this research sustains that association rules products with high OCVR values, are more susceptible of marketing strategies, as the determination and the dedication of the consumer in selecting the specific market basket presents significant variances. At the same time there should be observed the degree of confidence of rules with high OCVR and initially to control those who display a high degree of confidence. By taking action with appropriate marketing strategies it is more likely to change the level of confidence of those rules into a higher level, which already seems to be accomplished randomly and occasionally during the year. To change the location of the products on the shelf, to define the discount policy but also to form recommendation systems, the analysis of OCVR index values can be particularly useful. The evolution of consumer behavior in the course of time can reveal the combinations of products and, in particular, those products that are more suitable for promotion sales strategies.

6 Summary and Future Work

This research was conducted in Greek market product data and known super market chain, collected from customer purchases through loyalty cards. During the process of market basket analysis there were created 12 subsets of data corresponding to the data collected each month of the year. For each subset there were calculated lift and confidence indicators of association rules and at the same time a new indicator was co-estimated, the one of overall variability (OCVR). The OCVR results from the average of the values of the variability of lift and confidence indicators and describes degree of the overall variability of association rules. Each marketing strategy must be implemented to streamline the consumer behavior and to increase customer confidence. For this reason you must first identify application rules with high values of OCVR and high degree of confidence and attempt with the appropriate marketing strategy to reduce the value of OCVR while increasing the confidence of its customers.

In a future research it would be very interesting to establish an indicator which incorporates and includes the above information. It would also be quite interesting in the evaluation of OCVR to find weight factors for the representation of the variability of the degree of confidence and the degree of lift in the composition of the index. The existence of possible relations of substitution or even complementation between the association rules it may also constitute a subject for further research. Finally, it would be interesting to examine the existence of annual periodicity of the indexes, using additional data of more than one year.

Acknowledgments. We would like to thank Super Markets “Afroditi” for providing data and for their help.

References

1. Chen, Y.L., Tang, K., Shen, R.J., Hu, Y.H.: Market Basket Analysis in a Multiple Store Environment. Decision Support Systems 40, 339–354 (2005)
2. Shaw, M.J., Subramaniam, C., Tan, G.W., Welge, M.E.: Knowledge Management and Data Mining for Marketing. Decision Support Systems 31, 127–137 (2001)
3. Lawrence, R.D., Almasi, G.S., Kotlyar, V., Viveros, M.S., Duri, S.S.: Personalization of Supermarket Product Recommendations. Data Mining and Knowledge Discovery 5, 11–32 (2001)
4. Margaret, H.D.: Data Mining Introductory and Advanced Topics. Prentice Hall, Englewood Cliffs (2002)
5. Agrawal, R., Imielinski, T.: Mining Association Rules Between Sets of Items in Large Databases. In: Proc. of the ACM SIGMOD International Conference on Management of Data, Washington, D.C. (1993)
6. Ale, J.M., Rossi, G.H.: An Approach to Discovering Temporal Association Rules. In: Proc. of the 2000 ACM Symposium on Applied Computing, Como, Italy (2000)
7. Bayardo, R.J., Agrawal, R.: Mining the Most Interesting Rules. In: Proc. of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA (1999)
8. Wikipedia: The Free Encyclopedia, `http://en.wikipedia.org/wiki/Association_rule_learning`
9. Bhanu, D., Balasubramanie, P.: Predictive Modeling of Inter-Transaction Association Rules-A Business Perspective. International Journal of Computer Science & Applications 5, 57–69 (2008)
10. Brin, S., Motwani, R., Ulman, J.D., Tsur, S.: Dynamic Itemset Counting and Implication Rules for Market Basket Data. In: Proc. of the 1997 ACM-SIGMOD Conference on Management of Data, Arizona, USA (1997)
11. Anderson, C.: The Long Tail: Why The Future of Business Is Selling Less of More. Hyperion (2006)
12. Roiger, R.J., Geatz, M.W.: Data Mining: A Tutorial-Based Premier. Pearson Education, London (2003)
13. Bose, I., Mahapatra, R.K.: Business Data Mining – A Machine Learning Perspective. Information and Management 39, 211–225 (2001)
14. Chen, M.S., Han, J., Yu, P.S.: Data Mining: An Overview from a Database Perspective. IEEE Transactions on Knowledge and Data Engineering 8, 866–883 (1996)
15. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile (1994)
16. TANAGRA, a free DATA MINING software for academic and research purposes, `http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html`
17. Mild, A., Reutterer, T.: Collaborative Filtering Methods For Binary Market Basket Data Analysis. In: Liu, J., Yuen, P.C., Li, C.-H., Ng, J., Ishida, T. (eds.) AMT 2001. LNCS, vol. 2252, pp. 302–313. Springer, Heidelberg (2002)
18. Hong, S., Xinyan, Z., Shangping, D.: Mining Association Rules in Geographical Spatio-Temporal Data. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 196, 225–228 (2008)
19. Yo-Ping, H., Li-Jen, K., Sandnes, F.E.: Using Minimum Bounding Cube to Discover Valuable Salinity/Temperature Patterns from Ocean Science Data. In: Proc. of the 2006 IEEE International Conference on Systems, Man and Cybernetics, vol. 6, pp. 478–483 (2006)
20. Papadimitriou, J.: Statistical: Statistical Inference. Observer, Athens (1989)