Analyzing Price Data to Determine Positive and Negative Product Associations

Ayhan Demiriz[1], Ahmet Cihan[2], and Ufuk Kula[3]

[1] Dept. of Industrial Engineering
Sakarya University
54187, Sakarya, Turkey
ademiriz@gmail.com
[2] Dept. of Industrial Engineering
Kocaeli University
Kocaeli, Turkey
ahmet.can.cihan@gmail.com
[3] Dept. of Industrial Engineering
Sakarya University
54187, Sakarya, Turkey
ufukkula@gmail.com

Abstract. This paper presents a simple method for mining both positive and negative association rules in databases using singular value decomposition (SVD) and similarity measures. In literature, SVD is used for summarizing matrices. We use transaction-item price matrix to generate so called ratio rules in the literature. Transaction-item price matrix is formed by using the price data of corresponding items from the sales transactions. Ratio rules are generated by running SVD on transaction-item price matrix. We then use similarity measures on a subset of rules found by Pareto analysis to determine positive and negative associations. The proposed method can present the positive and negative associations with their strengths. We obtain subsequent results using cosine and correlation similarity measures.

1 Introduction

Data mining is used for discovering knowledge from large databases. As being an interdisciplinary approach data mining utilizes algorithms developed in computer science, mathematics, artificial intelligence, machine learning, statistics, optimization and other fields. As one of the early tools of recommender systems [1] Apriori algorithm [2] has been widely used for finding positive item associations. Apriori algorithm searches for the relations between product groups satisfying user supplied support and confidence levels and finds frequently bought product groups by customers. Although Apriori algorithm and its variations mostly use transactional data format, some forms of it require the data in transaction-item matrix format. Basically this type of matrix consists of binary data. Transaction-item (user-item) matrix is also the source of the data used in various collaborative filter based recommender systems.

C.S. Leung, M. Lee, and J.H. Chan (Eds.): ICONIP 2009, Part I, LNCS 5863, pp. 846–855, 2009.

Like the main stream research in association mining, item price data have long been neglected in recommender systems for finding relations between items. From this point of view, Apriori algorithm has a disadvantage of omitting the price paid by the customers to the purchased products despite of readily available data in transactions. This paper explores possibility of using transaction-item price data to find relationships (associations) between items. An early work, [3], studies ratio rules derived from the expense data to understand how much money would be spent on an item compared to the other items. In [3], a sample supermarket expense dataset was used in constructing the discussions for the ratio rules. Singular Value Decomposition (SVD) is used for finding the ratio rules which simply are eigenvectors corresponding to eigenvalues.

Similarly we adopt transaction-item price dataset from apparel retailing to assess the usability of price data for finding item relations. Our ultimate goal is to use these results in determining cross-price elasticities among multiple items. However our early findings indicate that we can use these results for determining both positive and negative item relationships as well. Our approach is summarized as follows: We first use SVD to decompose the transaction-item price matrix to find the eigenvectors i.e. ratio rules. We then deploy Pareto analysis to determine the important rules. This is indeed equivalent to picking the most influential eigenvalues and their eigenvectors. We then utilize some similarity measures, specifically cosine and correlation coefficient, to determine the sign and strength of relationships between items.

We also compare the outcome of our approach with traditional association mining results in this paper. We show that some of the positive associations can be recovered by our approach, however some associations are not found by our approach. This is indeed an important indication of the price sensitivity of the associations. Meaning that if the prices items high at the beginning, which is the case for the apparel retailing, items are more likely purchased alone. However the prices of the items are reduced as season progresses and as the prices of the items are marked down appropriately, it becomes more likely that certain items would be purchased together. This will obviously contribute a positive affect on the associations among such items.

Our aim in this paper is to show that item price data could potentially useful in determining positive and negative relationships between items. We summarize the contributions of the paper in the remaining of this section.

1.1 Contributions

The following contributions are provided in this paper:

- Transaction-item price matrix has been utilized in an association mining framework,
- Positive and negative relationships can be found by using transaction-item price matrix,
- Evidence is presented that positive associations can be attributed to the price reductions.

As listed above the paper has three main contributions. The rest of the paper is structured as follows. In Section 2, we give a brief description of the preliminaries. In Section 3, we introduce our methodology. A short illustrative example is presented in Section 4. We present results of our approach on a real dataset coming from apparel retailing in Section 5. We then conclude our paper in Section 6.

2 Preliminaries

Ratio rule mining technique uses eigensystem analysis. We can use SVD to find eigenvalues and eigenvectors of a non-square matrix. The number of eigenvalues of a matrix is equal to rank of this matrix. SVD method can simply be described for the matrix X, with transaction (customer) information in rows and product information in columns, by the following formula:

$$X = U \times \wedge \times T' \tag{1}$$

U and T are orthonormal matrices called left and right singular values respectively. $\wedge$ is the diagonal matrix with eigenvalues of X corresponding amplitude of eigenvectors described by T. All of the eigenvectors described by T are not used as ratio rules. There is a heuristic method for determining which eigenvalues are accepted for ratio rules [3]. According to this heuristic method the cutoff for the rules is %85 of the cumulative sum of eigenvalues. If the leading eigenvectors are very significant then using the rest of them as rules is unnecessary. Thus we can find a cutoff level for the rules by using Pareto analysis.

Pareto analysis is fundamentally using Pareto principle which can simply be phrased as follows: %80 of produced outputs are from %20 of inputs. To find which inputs have strong effects to generate the outputs you can plot the graph of inputs to corresponding outputs. This paper utilizes Pareto analysis as the number of eigenvectors in inputs and eigenvalues as outputs. The worst case scenario is that the eigenvalues are all equal. In this case, the Pareto plot has a slope of 45 degrees. For this reason, the cutoff level is determined by the slope of the line segments where the slope is lower than of 45 degrees. In other words if a line segment has a slope lower than 45 degrees it can be considered as the cutoff point for the rules.

3 The Methodology

Generating eigenvectors i.e. ratio rules is a straight forward step in our framework. After determining the most significant rules (i.e. truncated SVD) by Pareto analysis, we can deploy some similarity measures to summarize the relationships between products. In the literature there are many similarity measures [4] used for many different problem types.

The purpose of this paper is to find both positive and negative relationships (similarities) between products on significant rules found by SVD. There are two types of important information embedded in these rules. The first one summarizes the amount (i.e. ratio) of price paid, which represents the general behavior.

The second one is the sign information which represents the direction of the relationships between products. Therefore, we can deploy transaction-item price data to find the relationships between products.

One could use more measures to find similarities, however, for the brevity of the study we use two of them: correlation and cosine. Correlation and cosine similarity measures can vary between -1 and 1. If the value of the measure is negative, this means that the products have negative association between them. If the value of the measure is near zero then it can be concluded that the products are not related. Otherwise, if the value of the similarity measure is positive, then it can be concluded that the products have positive association between them. We give brief definitions of these similarity measures below.

3.1 Correlation Similarity Measure

Correlation coefficient similarity measure can be expressed by the following equality: $\rho(x, y) = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}$ where σ_{xy}^2 represents covariance between vector x and vector y and σ_x represents the standard deviation of vector x. Correlation coefficient is a widely used statistic in determining significant linear relationships.

3.2 Cosine Similarity Measure

Cosine similarity measure depends on the degree between two vectors. Cosine similarity measure can be expressed by equality: $\cos(x, y) = \frac{x \cdot y}{\|x\|\|y\|}$ where $x \cdot y$ is the dot product between vector x and vector y. $\|x\|$ is the 2-norm of vector x.

4 An Illustrative Example

In the following, we present an illustrative example (see Table 1) to depict our methodology. Each transaction is composed of price paid for three products. Notice that this is just a sample data. Obviously, if a product is not purchased at all then the corresponding price is equivalent to 0. We can potentially consider this dataset as an expense dataset, since the numbers correspond to the amount paid. However in our study we prefer to call it transaction-item price matrix. Table 1 lists the sample data used in this section.

After applying SVD to the data matrix, we find the eigenvalues and eigenvectors as follows:

$$\wedge = \begin{pmatrix} 11,6123 & 0 & 0 \\ 0 & 7,2180 & 0 \\ 0 & 0 & 3,1708 \end{pmatrix}$$

$$T' = \begin{pmatrix} -0,5036 & -0,6440 & -0,5759 \\ -0,6414 & -0,1678 & 0,7486 \\ 0,5788 & -0,7464 & 0,3286 \end{pmatrix}$$

Table 1. Illustrative Example Data

Transaction	Product 1	Product 2	Product 3
Transaction 1	3	1	0
Transaction 2	2	2	0
Transaction 3	2	1	0
Transaction 4	5	5	0
Transaction 5	0	1	4
Transaction 6	0	2	2
Transaction 7	0	1	2
Transaction 8	0	2	5
Transaction 9	0	3	1
Transaction 10	1	3	4
Transaction 11	4	2	3

Matrix T' corresponds to eigenvectors of matrix X and diagonal of matrix $\wedge$ corresponds eigenvalues of corresponding eigenvectors. In Figure 1, the cumulative importance of the rules derived from the eigenvalues is depicted against the number of rules considered. The slopes of the plot indicate the importance of the rules.

Eigenvalue that results in a line segment with a slope under an angle of 45 degrees is the cutoff for rules. However, in order to have a similarity measure we need at least two eigenvectors in our analysis. Since we have three eigenvectors (T'), we can only use two of them for a similarity measure. Notice that if we use all the eigenvectors (ratio rules) in our analysis then the similarity measures, for example cosine, will yield meaningless result that all the products are unrelated. This is due to the fact that all the eigenvectors are orthogonal to each other.

The sample case has the line segment slopes of [1.58 0.98 0.43] corresponding to Rule 1, Rule 2, and Rule 3, respectively. The first line segment has a slope bigger than 1. Technically we should avoid including Rule 2 to our analysis since it has a slope lower than 1. However we need at least two rules to generate similarity measures. The first two rules are given again below.

- Rule 1: [-0.5036, -0.6440, -0.5759]
- Rule 2: [-0,6414, -0.1678, 0.7486]

Based on the above rules we will have the following matrix which can also be called as ratio rules matrix (RR) to determine similarities:

$$RR = \begin{pmatrix} -0.5036 & -0.6440 & -0.5759 \\ -0,6414 & -0.1678 & 0.7486 \end{pmatrix}$$

The columns of the ratio rules matrix above correspond to the products. Similarity measures can be calculated by using this matrix to determine product relations (similarities). Using similarity measures over columns of ratio rules (rule-product) matrix results in product to product similarities. Notice that we

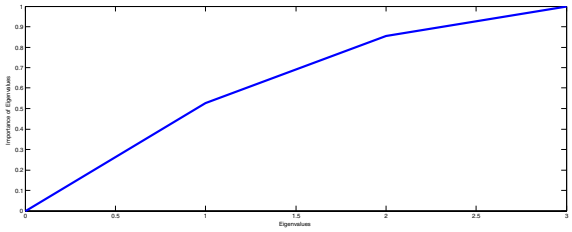

Fig. 1. Pareto Graph for Toy Example

can also use the sign of each value of the ratio rules matrix for determining product similarities. In this case we will have the following discretized ratio rules matrix to determine product similarities.

$$\begin{pmatrix} -1 & -1 & -1 \\ -1 & -1 & 1 \end{pmatrix}$$

After applying cosine based similarity measure on ratio rules matrix RR above, we get the following product similarities for the sample problem:

$$\begin{pmatrix} 1 & 0,7959 & -0,2469 \\ 0,7959 & 1 & 0,3901 \\ -0,2469 & 0,3901 & 1 \end{pmatrix}$$

A correlation coefficient measure will be as follows:

$$\begin{pmatrix} 1 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & 1 \end{pmatrix}$$

If we use a discretized ratio rules matrix on cosine based similarity measure, this will yield the following product similarities:

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Similarity measures over continuous data shows that product 1 and product 3 have negative association between them. Applying the correlation coefficient similarity measure on discrete rules yields inconclusive results. Similarity measures over discrete rules are inconsistent, because there is no variation (univariate) in other words the standard deviation is equal to 0.

5 Analysis

We use the sales data of summer season of year 2007 from a leading apparel retail firm in Turkey for the analysis. Like in any other retail environment, the

Merch. Group ⟶ Category ⟶ Model ⟶ SKU

Fig. 2. A Typical Representation of the Product Hierarchy (shown horizontally for brevity)

products in an apparel retail firm can be represented in a hierarchy. A typical hierarchy is shown in Figure 2. The major layers are shown in this hierarchy where merchandise group is shown at the top of the hierarchy, however additional layers can be inserted depending on the structure of the apparel business. Stock keeping unit (SKU) layer is the lowest level in this hierarchy. However SKU level data include unnecessary detail for the analysis. So we decided to use the model level data i.e. the data is aggregated at the size and the color levels for a particular garment.

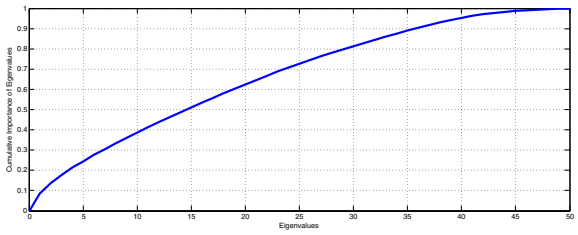

Fig. 3. Pareto Graph of Eigenvalues

For the purpose of this study, we pick the top 50 models out of 710 models belonging to one merchandise group based on the sales figures. Since the firm has provided us data belonging to a particular merchandise group (e.g. women's apparel), the top 50 models are from the same merchandise group. We then select the transactions with at least 5 products (items) involved (purchased) to reduce the adverse effect of sparsity of the data matrix. This gives us 3,525 transactions from the sales data with 50 models i.e. a 3525×50 data matrix.

There are 50 eigenvalues and corresponding eigenvectors found by using SVD. By applying Pareto analysis and visually inspecting the Figure 3, it is acceptable to conclude that approximately the leading 30 eigenvalues are significant for the given data matrix. We can then calculate the similarity measures to determine the product relationships. It should be noted that the similarity measures used in this paper vary between -1 and 1. In such a scale, measure values near zero (in both directions) represent unrelated products. However there is no clear cut threshold to determine the separation. The lower threshold (nearer zero) is, the more relationships will be found from the similarity measure matrix. For example, in Figure 4, we vary the threshold for the negative relationships between -0.1 and -1. In other words, if we have a similarity value between two products lower than the threshold level (since the similarity in the negative side of the spectrum), we can conclude that these two products have a negative

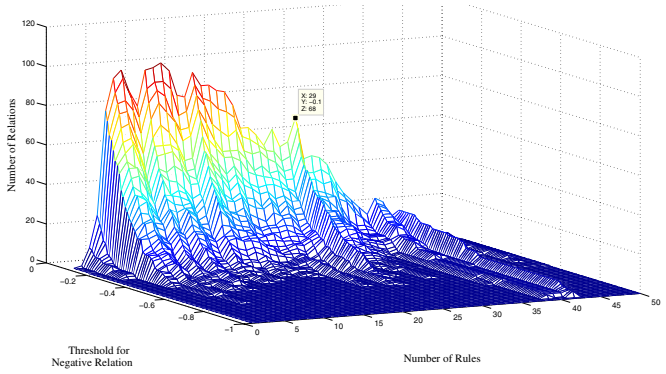

Fig. 4. Number of Negative Relations based on Cosine Similarity

relationship. In the same figure, it is evident that by using the first 29 ratio rules (eigenvectors) with a -0.1 threshold level from cosine similarity measure, we will find 68 negative relationships among 50 products. Correlation coefficient based similarity measure yields 75 negative relationships which has comparable number of relationships with cosine similarity. Out of 68 negative relationships found by the cosine similarity measure, there are only four relationships that could not be accounted by the correlation similarity measure. We can conclude that both measures behave similarly in terms of finding negative relationships. By using the discretized ratio rules, we usually find more relationships than continuous case in our experiments. However these relationships are questionable as seen in the illustrative example given in Section 4.

Similarly, we can vary the similarity threshold to observe the positive relationships as in Figure 5. Recall that a similarity threshold means that any two items which have a positive similarity measure above this threshold are considered similar. For the positive relationships, at 0.1 threshold level cosine similarity measure finds 168 positive relationships by using 29 ratio rules mentioned above. Based on the correlation coefficient similarity measure, our approach finds 177 and 75 positive and negative relationships respectively. Again we use 0.1 and -0.1 threshold levels for the positive and negative relationships respectively. Out of 168 positive relationships found by cosine similarity measure, there are only three relationships that could not be found by the correlation similarity measure which covers 177 positive relationships. These three relations that are not accounted by the correlation similarity measure are the borderline cases i.e. they are just below the threshold level. Again, we can conclude that both cosine and correlation similarity measures behave similarly in terms of finding positive relationships as well.

To compare our approach with the traditional association mining, we apply Apriori algorithm with a support count level 100 which is approximately 2.84% support and 10% confidence levels. We find 73 frequent pairs i.e. positive relationships meaningful. There are 24 pairs overlapping with our approach (cosine

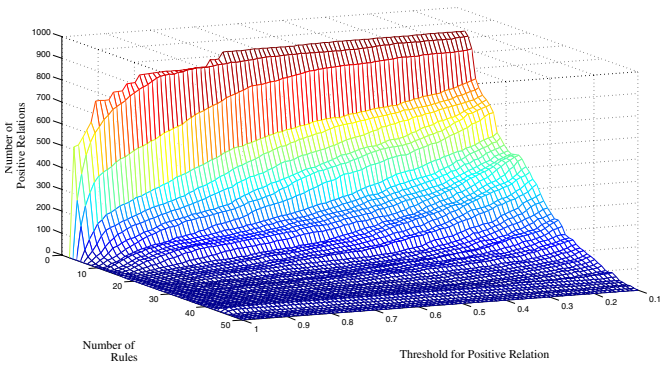

Fig. 5. Number of Positive Relations based on Cosine Similarity

similarity measure at 0.1 threshold level with 168 positive relationships) out of 73 frequent pairs from Apriori algorithm. For the negative association, we utilize indirect association mining from [5] which yields 91 negative relationships. Very few of 68 relationships found by our approach match with the results from indirect association mining.

Discrepancies between the traditional association mining and our approach can be attributed to the price sensitivities (multiple items cross-price elasticities) of the products. In apparel retailing, the price of the items are always higher at the beginning of the season. Later in the season, there might be significant reductions in the prices. When the prices of two items are sufficiently lowered, then the likelihood of purchasing both items increases. If both items are purchased together in a significant level during the sales season, the traditional association mining can pick this behavior as a positive association. However both items can show a different behavior at normal price levels. That's why our approach can identify this relationship as negative, since both items are not usually purchased together at normal prices, but at highly reduced prices.

6 Discussion and Conclusion

We have shown that transaction-item price data can be utilized for finding both positive and negative relationships. We also compare our approach with traditional association mining techniques: Apriori and indirect association mining.

Our analysis indicate that it may not always safe to conclude from a traditional association mining that two items have positive association for all the time, even though they satisfy the minimum support and confidence level constraints. This conclusion might be true if only both items are on sale at significant price reductions. In addition, we should point that the behavior of Apriori algorithm might change drastically at different price levels. To our best knowledge, there are no published results pointing this issue before.

Acknowledgement. This study is supported by the Turkish Scientific Research Council through the grant TUBITAK 107M257.

References

1. Demiriz, A.: Enhancing product recommender systems on sparse binary data. Journal of Data Mining and Knowledge Discovery 9(2), 147–170 (2004)
2. Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases. In: SIGMOD Conference, pp. 207–216 (1993)
3. Korn, F., Labrinidis, A., Kotidis, Y., Faloutsos, C.: Quantifiable data mining using ratio rules. VLDB J. 8(3-4), 254–266 (2000)
4. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, July 23-26, 2002, pp. 32–41 (2002)
5. Tan, P.N., Kumar, V., Kuno, H.: Using sas for mining indirect associations in data. In: Western Users of SAS Software Conference (2001)