Ertek, G., Tunç, M.M., (2012) "Re-Mining Association Mining Results through Visualization, Data Envelopment Analysis, and Decision Trees", in Computational Intelligence Applications in Industrial Engineering, Ed: C. Kahraman, Springer.

RE-MINING ASSOCIATION MINING RESULTS THROUGH VISUALIZATION, DATA ENVELOPMENT ANALYSIS, AND DECISION TREES

Gurdal Ertek

Faculty of Engineering and Natural Sciences, Sabanci University
Sabanci University, 34956, Orhanli, Tuzla, Istanbul, Turkey

Murat Mustafa Tunc
Faculty of Engineering and Natural Sciences, Sabanci University
Sabanci University, 34956, Orhanli, Tuzla, Istanbul, Turkey

Abstract

Re-mining is a general framework which suggests the execution of additional data mining steps based on the results of an original data mining process. This study investigates the multi-faceted re-mining of association mining results, develops and presents a practical methodology, and shows the applicability of the developed methodology through real world data. The methodology suggests re-mining using data visualization, data envelopment analysis, and decision trees. Six hypotheses, regarding how re-mining can be carried out on association mining results, are answered in the case study through empirical analysis.

1. Introduction

This study investigates how the results of association mining analysis can be analyzed further using data visualization, data envelopment analysis, and the computational intelligence method of decision trees.

Association mining is a popular data mining technique for operations management and industrial engineering, and reveals associations between item sets within a large item set[1]. While extensive literature exists on how to carry out association mining in a more efficient way, research on the interpretation of association mining results is relatively narrow. The initial motivation for data mining and association mining, which is the discovery of actionable insights, has been mostly neglected in theoretical and algorithmic studies. What is therefore needed is a set of methodologies that enable further analysis of data mining results, including association mining results that will yield further discoveries.

Re-mining is a general framework which suggests the execution of additional data mining steps based on the results of a data mining analysis[2]. Re-mining extends post-mining: Post-mining directly analyzes the data mining results, whereas re-mining suggests the incorporation of new attributes into the initial data mining results, enabling richer analysis and deeper insights. Even though re-mining suggests a general approach, there is considerable opportunity with respect to how it can be executed for various domains, data structures, and data mining goals.

This chapter suggests three types of analysis for re-mining of association mining outputs: As a pre-processing step, graph theoretic metrics are computed. Then, as the first analysis, graph visualization is used for the display of items and their associations and to discover actionable insights. Secondly, Data Envelopment Analysis (DEA) is adopted to formalize the

insights from the first analysis, and to identify items that deserve most focus in planning. Finally, decision tree analysis is carried out to obtain additional insights regarding how domain attributes and graph metrics affect the tendency to get involved in only positive associations.

Ceteris paribus, given that every other factor remains the same, positive associations are highly soughtafter in business practice. A manager would be highly interested in not only identifying the positive and negative associations in items, but also the percentages of the positive and negative associations for an item. A frequent item that engages in positive associations can be considered as an attractor item, which increases the sales of the items it clusters with [3]. This is the reason that the presented research focuses on the tendency of an item to get involved in only positive associations.

The contribution of this chapter is three-folds: Firstly, this is the first study, to the best of our knowledge, that analyzes association mining results using graph theoretical metrics. Our study formalizes how this analysis can be carried out through visualization and machine learning. Secondly, data envelopment analysis (DEA) is adopted for the first time in literature as a part of re-mining. Thirdly, the applicability and the usefulness of the proposed methodology is demonstrated through real world data coming from a case study.

The research questions explored in this study are as follows:

Question Q1. How can visual re-mining that considers both positive and negative associations of an item be carried out visually?

Question Q2. How can DEA be incorporated into re-mining?

Question Q3. Can the tendency of an item to engage in only positive associations be related to its domain-specific attributes?

Question Q4. Can the tendency of an item to engage in only positive associations be related to its metric values computed from its association graphs?

Question Q5. Can the relation in Q3 be predicted?

Question Q6. Can the relation in Q4 be predicted?

The remainder of the paper is organized as follows. Section 1.2 provides a brief review of some relevant literature as the background for conceptual development. Section 1.3 discusses the methodology developed. Section 1.4 is devoted to the demonstration of the

methodology through a case study with data from the real world. Finally, Section 1.5 draws some conclusive remarks from the research.

2. Literature

2.1. Association Mining

Association mining is a highly popular and useful data mining technique that is used in both academia and industry for both production and service systems[4]. The input to association mining is a transaction data, where each transaction contains a subset of items coming from a given set. The typical output of association mining, also referred to as association mining results, is the list of item sets that appear together frequently in transactions, and the rules that describe how these associations affect each other[1]. The first output is referred to as frequent itemsets, and the second output is referred to as association rules. The frequent itemsets and the association rules can be extremely numerous, and hence it is typical to specify a threshold support value while computing these results.

Support of an itemset (e.g.: {A,B}) or a rule (e.g. A⇒B) is the percentage of transactions that the items in the itemset or the rule appear in. Support is also the primary metric related with an itemset or rule that signifies importance. Another common metric in association mining is confidence, and is defined only for association rules (when only the above two outputs are generated). Confidence of a rule A⇒B is the conditional probability of item B appearing in a transaction, given that item A readily appears in that transaction. A mathematical description of the aforementioned concepts can be found in Demiriz et al[2].

The standard algorithm for association mining is the Apriori algorithm, which was first introduced by Agrawal et al.[5] and has gained increased popularity ever since (1991 citations for the original paper as of February 2012).Association mining is a standard module in almost every data mining platform[6, 7, 8] and can also be conducted through specialized software[9, 10].

The term association mining is commonly used to refer to only positive association mining. However, negative association rules are also essential in real world applications. Negative association does not simply refer to two items not appearing together. It refers to the situation of two items not appearing together in transactions as frequently as they should, when they are expected to do so, due to their independent associations with another mediator item. Two applicable methods for computing negative associations are introduced

in Savasereet al.[11]and Tan et al.[12], and the latter is applied in our study due to its convenience of implementation.

2.2. Graph Visualization

Graph drawing refers to the drawing of graphs, which consist of nodes and arcs, through specialized algorithms so as to obtain actionable insights[13, 14]. Graph visualization is a sub-field of information visualization[15, 16, 17]. Many successful applications of graph visualization are listed in Ertek et al.[4].

2.3. Association Graphs

In this chapter, the associations between items are represented as graphs, and are visualized using *grid layout* algorithm. Since the graphs presented in this study exclusively depict associations, rather than anything else, they will be referred to as association graphs[3, 4], following the terminology in Erteket al.[3]. In the presented association graphs, items are represented as nodes and the positive or negative associations between the items are represented as arcs. Grid layout algorithm snaps the items that are laid out according to a *force-directed algorithm* onto a grid, such that associated items are positioned closer to each other on the grid. The knowledge discovery from the grid layouts is enhanced through the mapping of additional attributes to the node colors and sizes.

2.4. Re-mining

Re-mining is a broadly applicable concept, which refers to the “mining of a newly formed data that is constructed upon the results of an original data mining process”[2]. In re-mining, the enriched data includes not only the results of the original data mining process, but also additional attributes. The goal is to obtain new insights that couldn’t have been discovered otherwise, and to characterize, describe, and explain the results of the original data mining process. Re-mining is a generalized extension of *post-mining*, which only summarizes the results of the original process. Re-mining can involve any type of data mining analysis, including *exploratory*, *descriptive*, and *predictive* analysis. In Demiriz et al.[2], the original data mining process is association mining, and it is proven through complexity analysis that the running time for re-mining is polynomial in problem size, whereas it is exponential in the alternative quantitative association mining (QAM).

2.5. *Re-mining on Association Graphs*

While the concept of re-mining was first introduced in a journal paper by Demiriz et al.[2], Ertek et al.[3] is the first study that formalizes *visual re-mining* on association graphs. Erteket al.[18]suggest a visual data re-mining process, based on association graphs, where values of additional attributes are mapped onto node colors. The authors demonstrate the applicability of the approach through a *market basket analysis* (*MBA*) case study. Specifically, the mentioned study maps statistics on the attributes of customers (such as gender, knowledge of French language, and hunger level) linearly to the node colors. The statistics displayed (through node color) within a node are computed only for the customers that select the item/itemset represented in that node. It is demonstrated that visual re-mining of association graphs enables the discovery of links between transactional patterns and customer attributes.

In this chapter, the node colors and sizes are not related to customer attributes, but are rather related to domain-specific item attributes and the tendency of the item to get involved in positive associations, rather than negative ones.

The visual re-mining on the association graphs answers research question Q1.

2.6. *Data Envelopment Analysis (DEA)*

Data Envelopment Analysis (*DEA*) [19] is a highly popular[20] and effective methodology that can be implemented to benchmark a group of entities through *efficiency scores*. Through the results of DEA, managers can obtain a multi-dimensional benchmark of the entities they are comparing, such as items, stores, suppliers, etc., and gain actionable insights on how to improve inefficient entities. These entities being compared are referred to as *DMU*s (Decision Making Units). In DEA, calculation of efficiency scores (the primary benchmark metric) for the DMUs is based on their input and output values. Inputs are the resources consumed by the DMUs, for generating desired outputs. Efficiency score, which takes a value between 0 and 1, increases as a DMU generates higher values of any of the outputs when the input values are the same, or when a DMU generates the same values of the outputs when any of the input values is lower.

Our methodology suggests the use of DEA to integrate the results from multiple visual analyses. In the presented case study, a DEA model was constructed to identify the most “efficient” items, where efficiency was defined as appearing in many transactions, engaging in mostly positive associations, and having a high price (specifically, a high end-of-season price), within a short lifetime.

The DEA model, which constitutes part of our methodology, answers research question Q2.

2.7. Graph Metrics

Graph theory is the field of mathematics that investigates problems related to graphs. A *graph* is composed of two types of entities, *nodes* that represent discrete, distinct entities and *arcs* that connect the nodes. The nodes in a graph and the graph itself can be characterized through *graph metrics*. In this chapter, several novel research questions are pursued as listed in introduction, where Q4 relates to graph theory. To this end, graph metrics are computed for each item (node on the association graph), and a classification model is used for prediction. The computations are done on two different association graphs, the one that displays only positive associations, and the one that displays only negative ones. The graph metric attributes are named with the suffix POS or NEG, based on which association graph the metric is computed for.

The graph metrics computed and used in this chapter are given as follows:

• *Degree* shows the number of connections for each node. It is an integer value, and it is the summation of in degrees and out degrees of the node.

• *Betweenness centrality* represents total number of shortest paths for each pair of nodes, if the node is on that path. It can take values between 0 and 1.

• *Closeness centrality* shows the distance between the node and every other node. It only takes values between 0 and 1.

• *Eigenvector centrality* shows the distance between the node and every other “special” node. It is a value between 0 and 1.

• *Page rank* is the value which only increases if the node is closely related with “special” nodes, in other words, its closeness-centrality of “special” nodes is higher[21].

• *Clustering coefficient* represents the tendency of aggregation for several nodes[22]. It can only take values between 0 and 1.

Detailed information about these and other graph metrics are explained in Opsahlet al.[23]and Christensen and Reka[24].

2.8. Decision Trees

*Decision trees*summarizerule-based information regarding classification as trees. In decision tree models, each node is split (branched) according to a criterion. Then, a tree is constructed with depth. At each level, the attribute that creates the most increase compared to the previous level is observed. The algorithms for decision tree analysis are explained in Chien and Chen[25]. In our study, ID3 algorithm[26] is used, and branches are created in Orange software[27]. In decision trees, identifying the nodes that differs considerably from its root node are our main focus. By observing the shares of slices and comparing with the parent nodes, one can discover classification rules.

The decision trees answer research questions Q3 and Q4.

2.9. Classification

In the classification models, the dataset is divided into two groups, namely *learning dataset* and *test dataset*. Algorithms that are referred as *classifiers* (or *learners*) use the learning dataset in order to learn from data and predict the class attributes in the test dataset. The prediction success of each learner is measured through*classification accuracy*(CA), which is the percentage of correct predictions among all[1]. The following classification algorithms are among the most well-known classifiers in the machine learning field: Naive Bayes, k-Nearest Neighbor (kNN), C4.5, Support Vector Machines (SVM), and Decision Trees[28].

The classification model searches for answersto two of the novel research questions posed in this chapter: Q5 and Q6.

3. Methodology

The methodologies described in the literature reviewed in section 2 have already been developed and shown to be applicable. Association mining, classification, and decision trees are especially well established methodologies within data mining[1, 29], and graph theory is a very extensive field of mathematics that has found its applications in almost every field of science[30, 31,32].DEA is a well-researched and vastly applied methodology for benchmarking[20]. Association graphs, re-mining, and re-mining on association graphs are relatively much newer methodologies. However, to the best of our knowledge, there does not exist a methodology that links these diverse methodologies in a unified framework. Specifically, our research shows that insights can be obtained for items that engage in only positive association (and no negative associations) based on their association graph metrics or

domain-specific attributes. Our research also shows that exploratory and descriptive re-mining may be very successful, whereas predictive re-mining may not yield satisfactory results.

The methodology that we propose consists of the following steps:

Step 1. Perform positive association mining on the transactions database for obtaining frequent item pairs (2-itemsets).

Step 2. Find negatively associated item pairs, using the results of Step 1.

Step 3. For each item, compute the percentage of positive associations it gets involved in.

Step 4. Construct two association graphs; one that shows only positive associations, and the other that shows only negative associations. Items are represented as nodes, and associations (positive or negative) are represented as arcs.

Step 5. Compute the graph metrics for each node (each item) in each of the two association graphs.

Step 6. Construct the dataset for re-mining, where each row is an item (that got involved in at least one positive association), and the columns are support count (SupC), domain-specific item attributes (StartWeek, EndWeek, …), percentage of positive associations (PercOfPositiveAssoc), class labels that are defined based on percentage of positive associations (PositiveAssoc takes value of 1, if PercOfPositiveAssoc=1)and graph metrics obtained from Step 4 (DegreePOS, BetweennessCentralityPOS, …, DegreeNEG, BetweennessCentralityNEG, …).

Step 7. Apply grid layout for the association graph, and map domain-specific attributes to node color and size. Visually analyze the graphs for discovering actionable insights.

Step 8. Construct a DEA model, to combine the insights from multiple visual analysis, and to find the most important items.

Step 9. Construct a classification model, to predict the items that engage in only positive associations (items that have PercOfPositiveAssoc = 1). Construct decision trees, using the classification model of Step 8.

Step 10. Apply multiple classifiers in the classification model and evaluate classification accuracy.

In the next section, the described methodology is applied to a real world dataset coming from an apparel retail company.

4. Case Study

4.1. Problem Description

The company that we work with inour research projects is one of the largest apparel retail chains in Eurasia. Headquartered in Istanbul, Turkey, the company has more than 300 retail stores in Turkey alone, as well as more than 30 stores in more than 10 other countries. The data analyzed comes from the men's clothes line (one of the merchandise groups) and belongs to the 2007 summer season. Earlier research with the data proved the applicability of the re-mining methodology[2], and focused on the analysis of item pairs. The findings in this chapter relate to the items as single entities.

The selection of retail industry for the case study is highly relevant, because this is an industry where the business is centered on customer transactions as the revenue stream. Therefore, novel and applicable data mining and machine learning methodologies for association mining can be incorporated into the retail operations. Yet, there is a bigger picture: As of November 2010, the retail industry in U.S. alone exceeded $ 377.5 billion[33]. This is a huge industry, and even the smallest percentage improvement in operations can correspond to significant financial gains.

4.2. Analysis Process

As **Step 1** of the proposed methodology, positive association mining was carried out on the transaction database for obtaining frequent item pairs. The minimum support count was taken as 100, which means that only the item pairs that appeared in at least 100 transactions were considered. A total of 3930 such frequent item pairs have been identified, involving 538 items.

Next, in **Step 2**, negatively associated item pairs were identified based on the items that formed the frequent item pairs. A total of 2433 negatively associated item pairs were identified, involving 537 items of Step 1 (all items except one of them).

In **Step 3**, the percentage of positive associations each item gets engaged in (PercOfPositiveAssoc) was computed. For this, all the associations for that item were filtered and the percentage was computed as the number of positive associations divided by the total number of associations. The distribution of PercOfPositiveAssoc is shown in Figure 1.

In **Step 4**, positive and negative association graphs were constructed using the Fruchterman-

Reingold force-directed algorithm[34]. It was not possible to obtain any immediate insights from the association graphs.

In **Step 5**, graph metrics were computed for each node in the association graph, using NodeXL add-in[35, 36]for Microsoft Excel. The computed metrics were the ones listed earlier in the literature review.

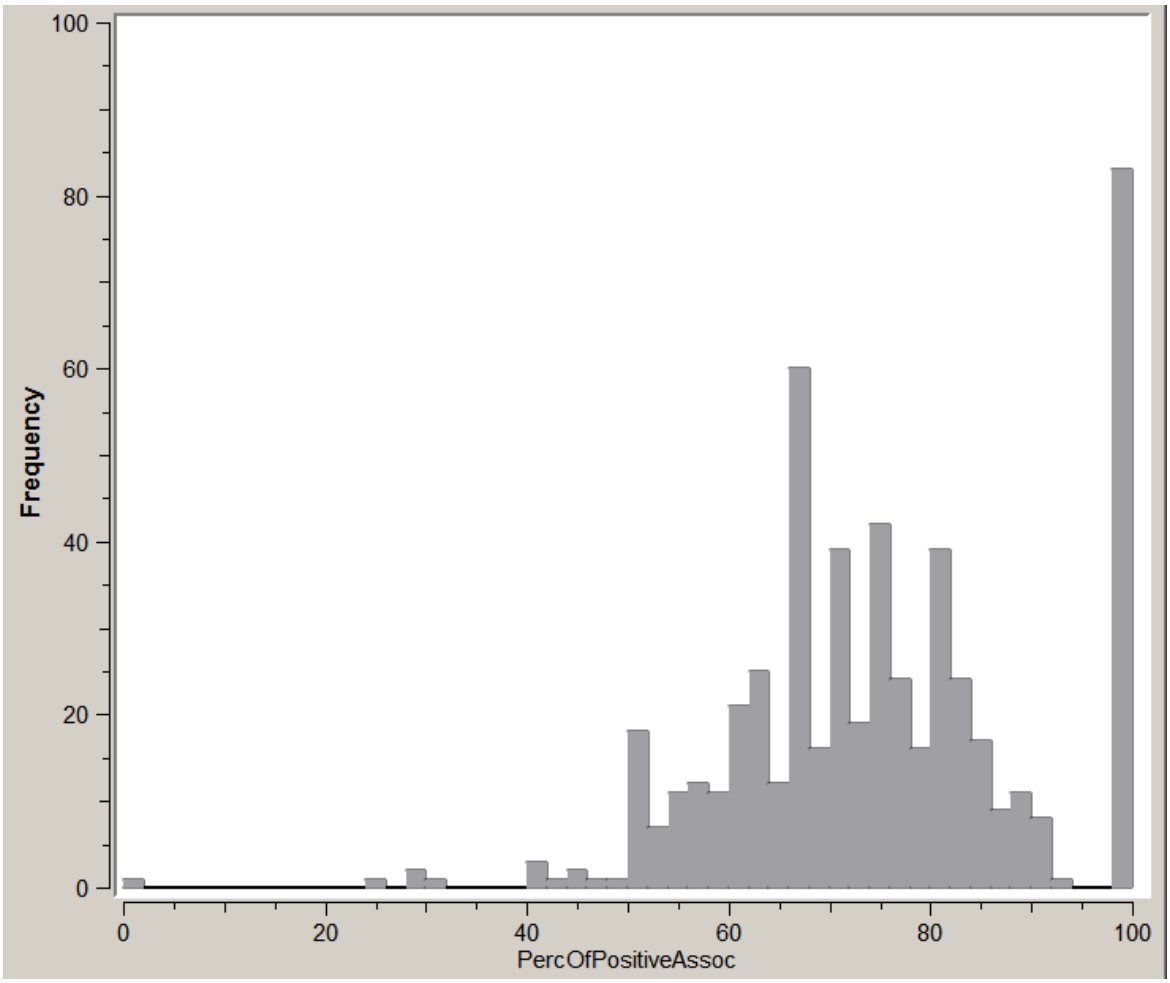

Fig. 1. The distribution of PercOfPositiveAssoc values for analyzed items.

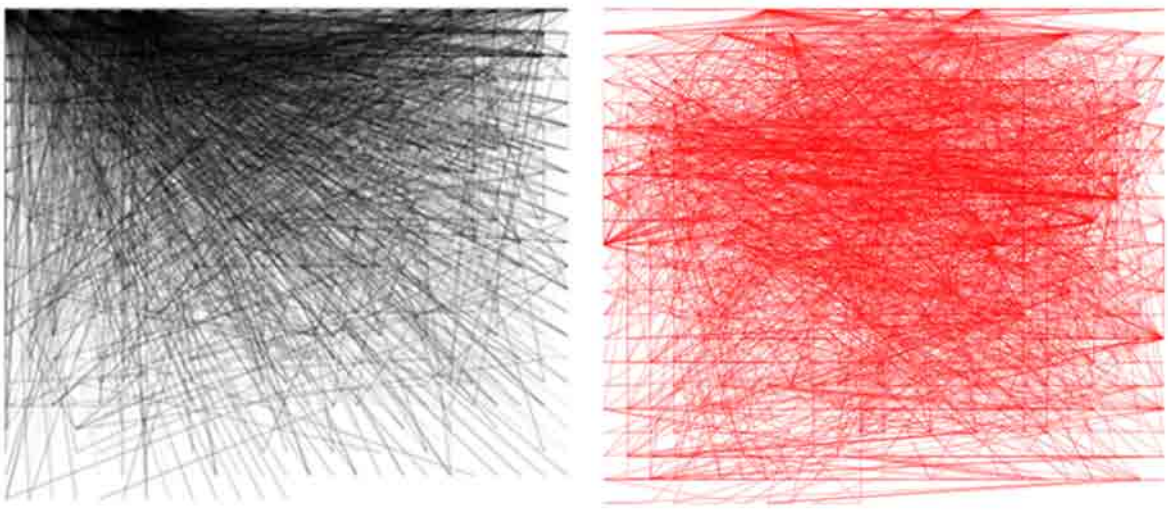

Fig. 2. The positive and negative association graphs for the transactions data set.

In **Step 6**, a dataset was formed for re-mining, where each row is an item that got involved in at least one positive association. The first column in the dataset (key attribute) is the unique item number. The second column in the dataset is the support count (SupC), which is the number of transactions the item appears in. The domain specific attributes are the following:

- StartWeek: The week within the season when the sale of that item was started.
- EndWeek: The week within the season when the sale of that item was ended.
- LifeTime: The number of week the item stayed on sale.
- MaxPrice: The initial price of the item (in apparel retail industry, item prices are subject to continuous markdowns, without any increases throughout the season).
- MinPrice: The final price of the item.
- PriceDiff: The difference between MaxPrice and MinPrice.
- MerchSubGrp: The merchandise subgroup the item belongs to.
- Category: The category the item belongs to.

The next column in the dataset is PercOfPositiveAssoc which was computed in Step 5. This continuous value was then discretized into a class attribute (PositiveAssoc), where the target class is AllPositiveAssoc (meaning that the item engages only in positive association, and no negative associations). The remaining columns are the graph metric values for the item.

In **Step 7**, grid layout was applied in NodeXL for visualizing the association graph, based on snapping the nodes positioned by the Fruchterman-Reingold algorithm[34] on a grid.The insights obtained in here are described detail later.

In **Step 8**, a DEA model was constructed using the SmartDEA Solver software[37], which implements the basic CCR and BCC models, and enables the generation of DEA outputs in a database format.

In**Step 9**, decision trees were constructed for descriptive re-mining using C4.5 algorithm (Figure 3).

Finally in **Step 10**, a classification model was constructed (Figure 3) in Orange data mining software[27, 38], to predict the class attribute PositiveAssoc. In our model for classification accuracy, learning dataset was taken from 70% of overall dataset, with 20 repetitive sampling. This means that for each learner, 20 random sampling of 70% of the dataset were executed as the learning dataset, and the remaining was experimented as the test dataset. In the classification models, random sampling was applied with %70 of the rows being the training set (376 rows), and the remaining %30 being the test set (162 rows).

In the classification step, three models, namely Model A, B, C, have been developed. In all three models, PositiveAssoc is the class attribute, in Model A, only domain related attributes are used as predictors; in Model B, only graph metric attributes are used as predictors; in Model C, both domain-specific attributes and graph metric attributes are used as predictors.

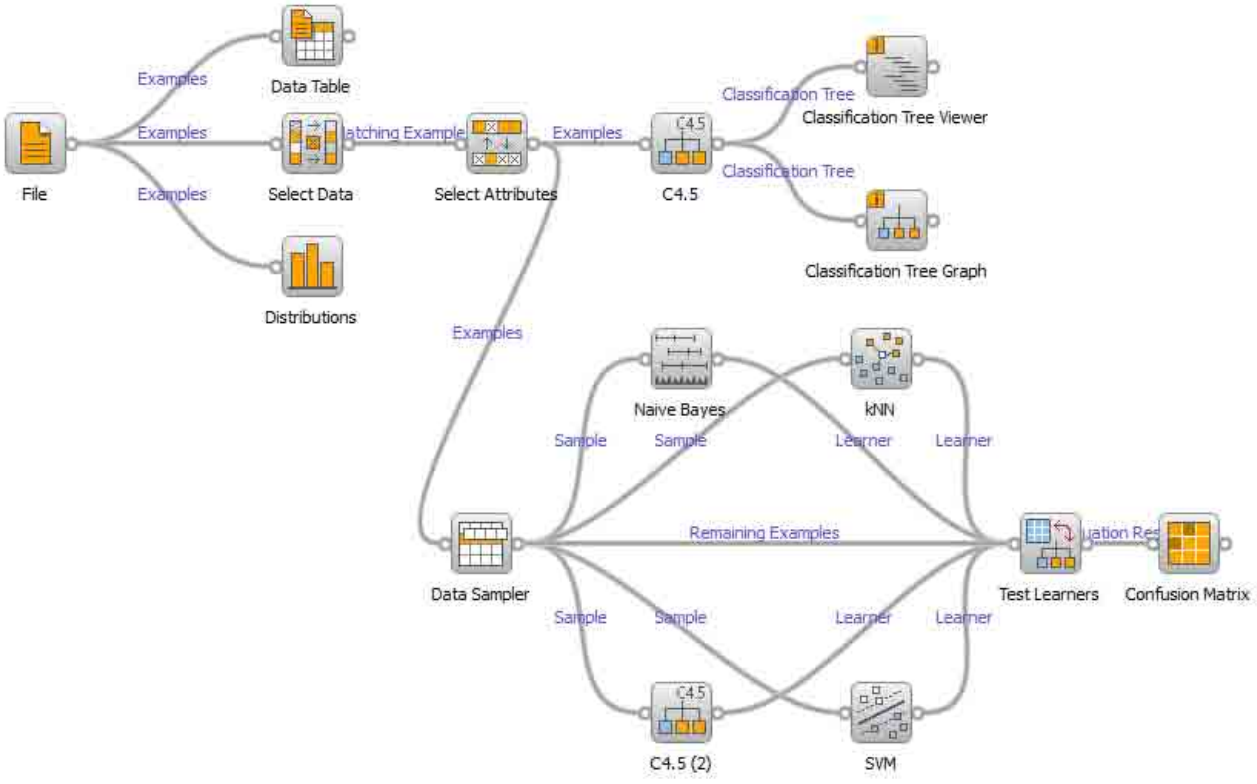

Fig. 3. The machine learning model for decision tree analysis and classification.

4.3. Graph Visualization

The analysis of association graphs can yield significant insights[4]. This section presents the analysis of association graph visualizations, where each node's color, size, and shape represents

attributes of that item. Actionable and practical insights and policies are obtained through the graph visualizations.

Each node in the visualizations shows a particular apparel item and each arc (connecting two nodes) shows an association (Figures2, 4, 5, 6). Black arcs denote positive associations and red arcs denote negative associations. The graph visualization is automatically constructed through the NodeXL graph analysis software, with the constraint of snapping all the nodes a rectangular grid. The color of a node can denote any selected variable. Since the focus of this research is the investigation of percentage of positive associations, color of each node is based on the PercOfPositiveAssoc value for that item. Yellow node color denotes items that engage in mostly negative associations, and darker nodes denote items that engage in more positive associations (higher PercOfPositiveAssoc values). Similarly, the size of can denote any selected variable. In our visualizations(Figure 5 and 6), the size of that node denotes the minimum price of that item.

The first observation in Figure4 is that the upper left quadrant of the graph consists mostly of nodes. A further examination of the node colors reveals that the node colors in general are lighter compared to the remaining regions of the graph, suggesting items that have many associations tend to get involved in more negative associations, as well. This is consistent with the definition and the computation method of negative associations.

Figure 4 also highlights very dark-colored outlier items within the items on the upper left region, which engage in not only many associations, but also mostly (or completely) positive ones. Everything else being the same, these items are very important, because they engage in many associations on the positive association graph, and do not engage in much negative associations. They complement many other items, and do not substitute many or even any items.

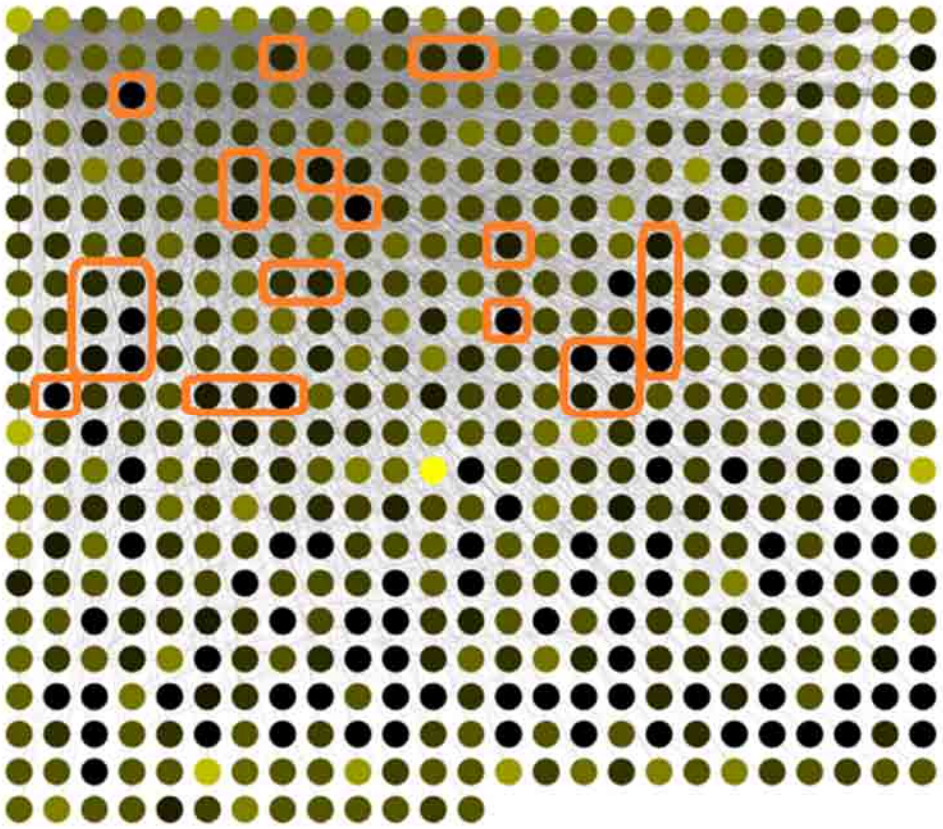

Fig. 4.Node color denotes the PercOfPositiveAssoc value of each item.

Figure 5 appends an additional dimension to Figure 4: The end-of-season sales prices (MinPrice) of the items, represented through the node sizes. Larger node sizes denote higher MaxPrice values. These high priced items also reside in the region that contains many positive associations, and are thus very important for the retailer. The suggestion for the retailer would be to understand the characteristics on these particular items and also sell these high-priced items throughout the season to maximize the revenue, both through their high-price and through the sales that they trigger as complements.

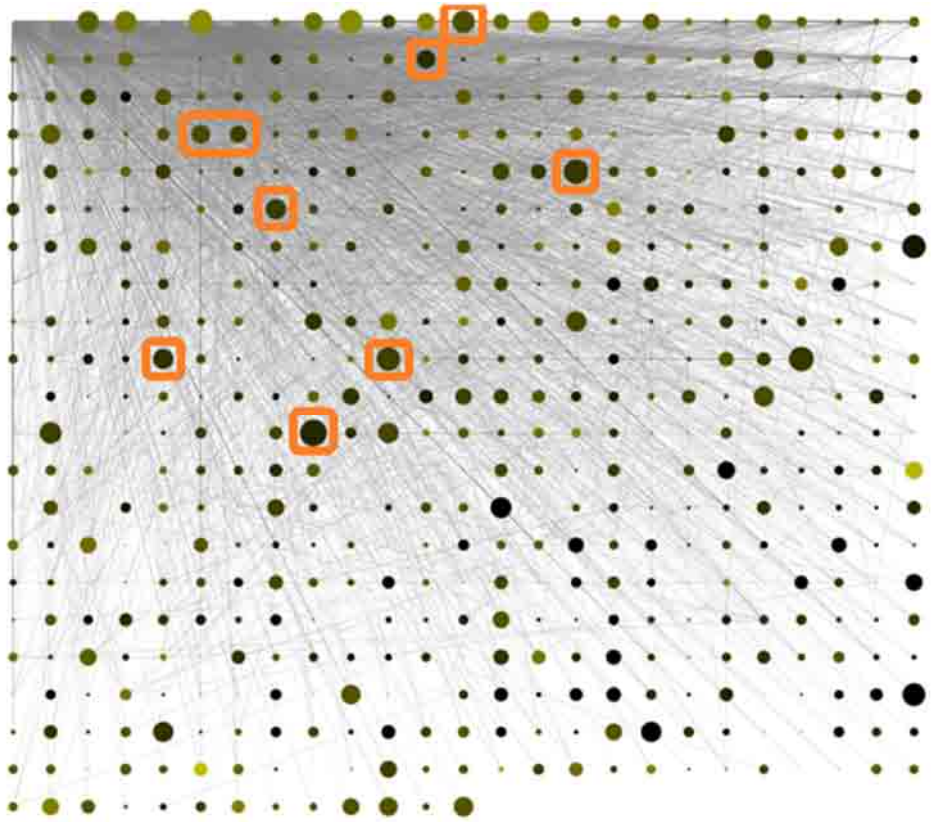

Fig. 5. Node size denotes the item's minimum price, and the node color denotes the item's PercOfPositiveAssoc value. The highlighted item triggers the sales of many items and also brings more revenue itself.

Finally, Figure 6 adds a new dimension to Figure 5: The categories of the items (category), represented through the node shapes. This analysis enables the identification of whether the important items in Figure 4 and 5 belong to a particular category or not. Highlighted nodes are in the upper left region of the positive association graph (involvement in positive associations with many items), have dark colors (high percentage of positive associations), have relatively large sizes (high end-of-season prices), and all come from only one of the three item categories (the category shown with a square). It is easy to see that almost all the items with the desired characteristics are from the same item category. The analysis in this section shows that exploratory visual re-mining brought significant insights and has thus been successful in this case study. The research question Q1 has been answered positively.

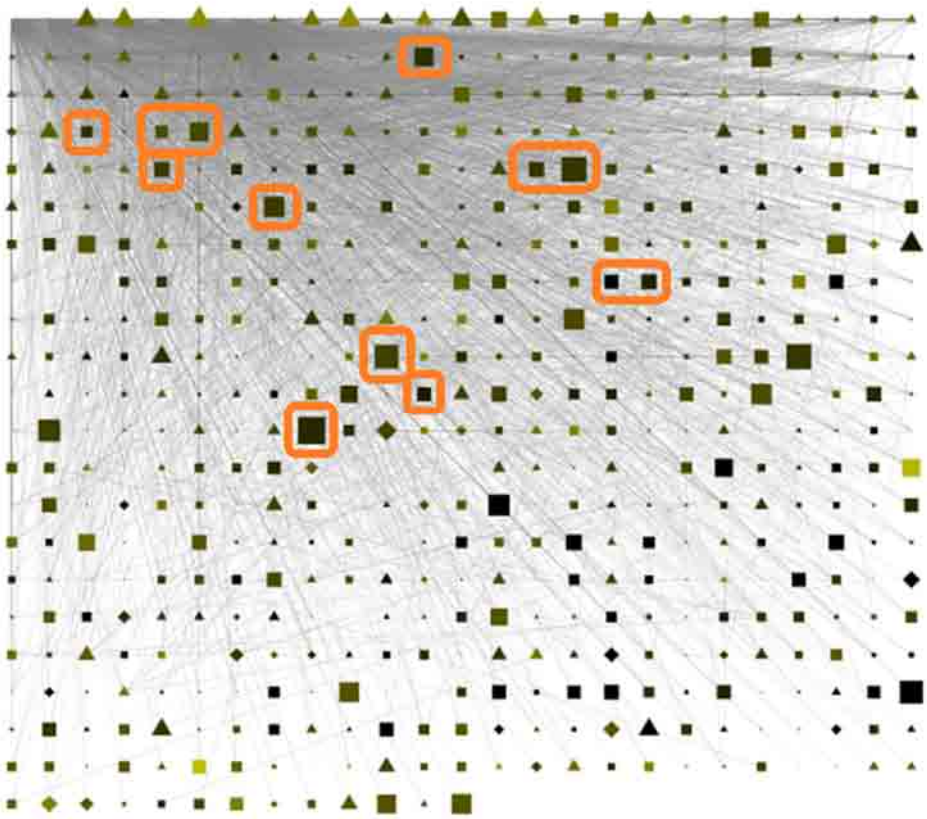

Fig. 6. The size of the nodes shows the minimum price of the items, the color of the nodes illustrates PercOfPositiveAssoc values, and node shape denotes the category.

4.4. Data Envelopment Analysis (DEA)

The DEA model was constructed so that the insights obtained from visual re-mining can be integrated within a unified analytical method. Thus, DEA is introduced not as a data mining method, but as a method to formally and analytically integrate the insights that can be obtained through the above visual data analysis. The DEA model considers three criteria, and treats their values as outputs. Since there were no inputs discussed in the visual re-mining section, there exists only a single auxiliary input, which takes the value of 1 for all the DMUs. The outputs are the support count (SupC), percentage of positive associations the items gets engaged in (PercOfPositiveAssoc), and the end-of-season price (MinPrice) of the items. Since the values for all the outputs are non-negative, the BCC models, which were introduced by Banker et al.[39] were found most appropriate selected, due to their allowance for variable returns to scale. Since the input values are auxiliary, all equal to 1, and cannot be changed, the goal of a DMU is to increase the values of its outputs (the values for performance criteria). This perspective is reflected in the selected DEA model, which is the output-oriented BCC model (BCC-O).

Table 1. Results of the DEA model, together with the inputs and outputs.

Item	Eff1*	Eff2**	Input_Auxiliary	Input_LifeTime	PercOfPositiveAssoc	SupC	MinPrice
059	Yes	Yes	1	16	91.67	4157	19.99
087	Yes	Yes	1	26	92.31	8947	14.90
094	Yes	Yes	1	11	75.00	4647	41.57
106	Yes	Yes	1	32	30.00	346933	9.25
169	No	Yes	1	7	75.00	4464	34.90
289	No	Yes	1	8	87.50	4317	23.06
412	Yes	Yes	1	13	88.89	2658	34.90
438	No	Yes	1	10	91.67	4999	14.90
513	No	Yes	1	4	80.00	5115	13.80

The results of DEA for the described model generate 84 efficient items. However, all but one of these items are efficient because they have PercOfPositiveAssoc=100, the highest value possible. Therefore, a new DEA model was constructed, that excluded these 83 items that engage in only positive associations. This time, five items were found to be efficient (first five items of Table 1). Finally, a new, third DEA model was constructed, that treated LifeTime as the only input. This

time, nine items were found to be efficient (Table 1). This third model considers the effect of life time of an item, and praises items that achieve more of the desired qualities in less time.The research question Q2 of how DEA can be integrated into re-mining is now answered.

4.5. Decision Trees

In both decision tree analysis (for Q3 and Q4) and classification analysis (for Q5 and Q6), three models, namely Model A, B, C, have been developed. In all three models, AllPositiveAssoc (whether the item engages only in positive associations) is the class attribute. In Model A, only domain related attributes are used as predictors; in Model B, only the graph metric attributes computed from the positive association graph are used as predictors; in Model C, only the graph metric attributes computed from the negative association graph are used as predictors. The list of predictors in each model is given below:

Model A: StartWeek, EndWeek, LifeTime, MaxPrice, MinPrice, PriceDiff;

Model B: DegreePOS, BetweennessCentralityPOS, ClosenessCentralityPOS, EigenvectorCentralityPOS, PageRankPOS, ClusteringCoefficientPOS

Model C: DegreeNEG, BetweennessCentralityNEG, ClosenessCentralityNEG, EigenvectorCentralityNEG, PageRankNEG, ClusteringCoefficientNEG

Figures 7-9 show the decision trees constructed for Models A, B, and C, respectively. In these decision trees, dark slices denote the share of items that engage in only positive associations (AllPositiveAssoc=AllPositive).

In Figure 7, it can be observed that items that have large price differences (>26.490) through the season have a higher chance of having only positive associations, as denoted by the larger share of the dark-colored slice. This chance increases if the item is taken out of the market before week 37, and increases even more if the item is introduced before week 22. These three pieces of information print out strategies on when an item should be introduced, until when it should be sold in stores, and how much of a price difference (PriceDiff=MaxPrice-MinPrice), it should have between its initial price (MaxPrice) and final end-of-season price (MinPrice).

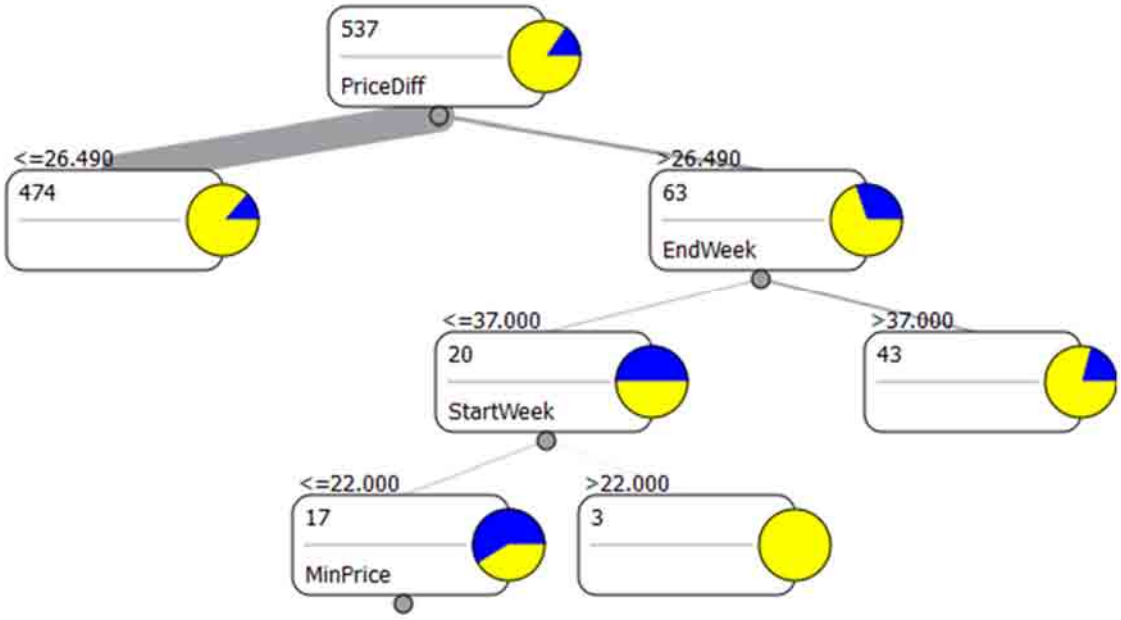

Fig. 7.Decision Tree for Model A.

In Figure 8, it can be observed that items that have a BetweenCentralityPOS=0 in the positive association graph have no chance of engaging in only positive associations. Among the other items, ClusteringCoefficientPOS ≤ 5 and even more when EigenVectorCentralityPOS ≥ 0.

How can the findings of this decision tree be used in practice? The results obtained from decision tree help us indentify the subgroup of items that have high chance of engaging in only positive associations. So, without carrying negative association mining for new data sets (not a readily available analysis type in many commercial software), one can obtain idea on how to search for important items.

Figure 9 shows the decision tree for Model C. This decision tree summarizes the behavior of all-positively associated items with respect to negative associations. Where would one find all-positively associated items the most? The answer is “when DegreeNEG=3 or 4 and ClusteringCoefficientNEG ≥ 0.4”.Research questions Q3 and Q4 have been answered in this section.

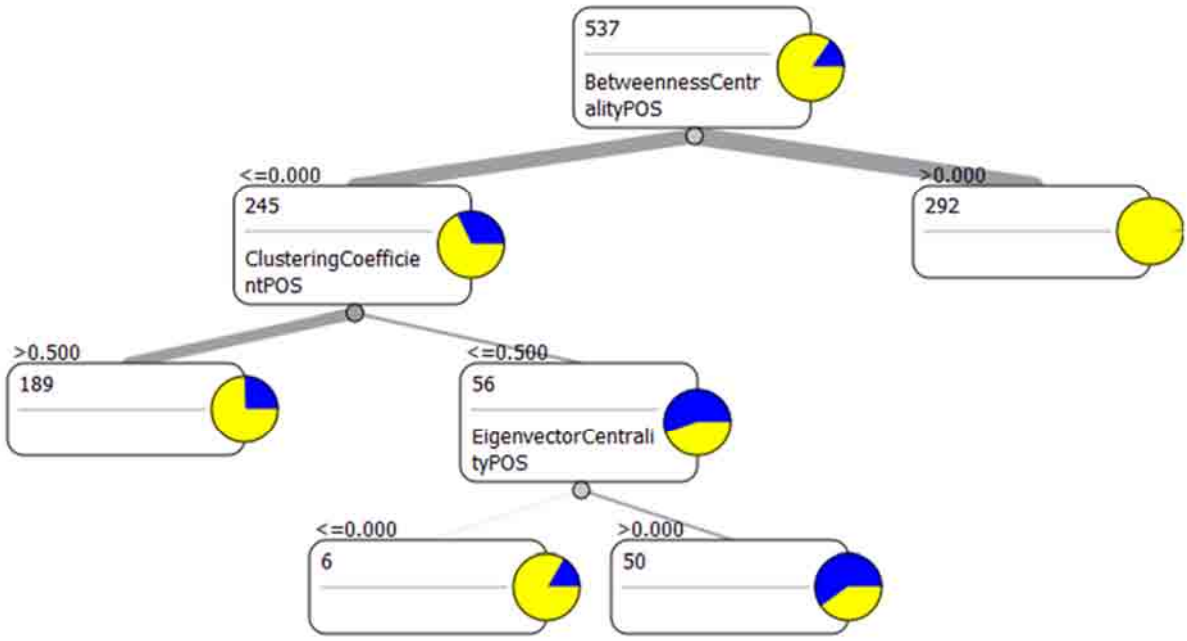

Fig. 8.Decision Tree for Model B.

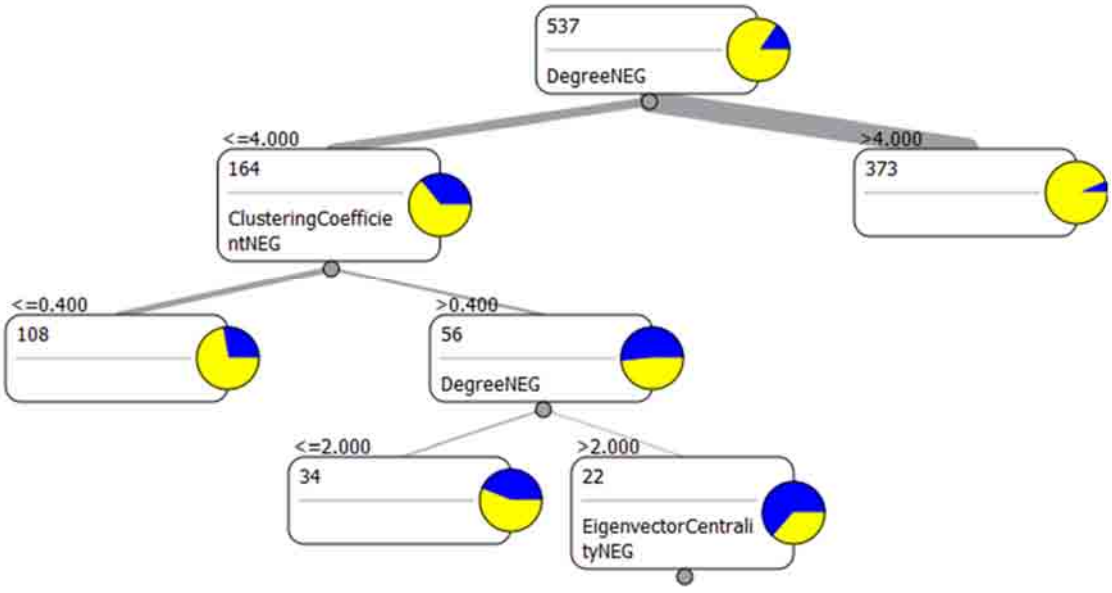

Fig. 9.Decision Tree for Model C.

4.6. Classification

The classification accuracies obtained for the three models using various learners are presented in Table 2. In all three models, it is possible to achieve a classification accuracy of at most 83.57%. However, the percentage of items with only positive associations is 15.45%, meaning that if a classifier guessed none of the items to have all-positive associations, the average classifier accuracy would still be 84.55%. Clearly, the classification models predict worse than a blind guess, and thus predictive re-mining through classification was not successful in this case study, despite the

success of exploratory and descriptive re-mining. Furthermore, a stepwise regression analysis, which is not described here, yielded a very low adjusted R2 value of 0.2180, suggesting that neither the domain attributes nor the graph metrics cannot be used in a linear regression model to predict PercOfPositiveAssoc. The answers to both Q5 and Q6 are “No”.

Table 2. Classification Results (Classification Accuracies).

Classifier	Model A	Model B	Model C
C4.5	0.8204	0.8000	0.7562
SVM	0.8357	0.8063	0.8094
kNN	0.7592	0.7406	0.7094
Naïve Bayes	0.7622	0.6313	0.7281

5. Conclusion

In this chapter, a novel computational intelligence methodology was introduced for practical re-mining. The methodology combines association mining, graph theory, classification, DEA, and re-mining, for answering six novel research questions.

The application of the methodology in a case study demonstrates the applicability and usefulness of exploratory and descriptive re-mining. The tendency of an item to engage in positive associations is related to its graph metric values computed from the association graphs, and to its domain-specific attributes. However, predictive re-mining through classification and stepwise regression did not prove useful in the context of the case study.

Acknowledgement

The authors thank Ayhan Demiriz of Sakarya University for providing the data for this study. The authors also thank Ilhan Karabulut for her work that inspiredthe visual re-mining approach on association graphs.

References

J. Han, M. Kamber and J. Pei, *Data Mining: concepts and techniques* (2011).

A. Demiriz, G. Ertek, T. Atan and U. Kula, Re-mining item associations: Methodology and a case study in apparel retailing, *Decision Support Systems*, 52(1), pp. 284-293. (2011).

G. Ertek and A. Demiriz, A framework for visualizing association mining results, *Lecture Notes in Computer Science (LNCS)*, 4263, pp. 593-602. (2006)

G. Ertek, M. Kaya, C.Kefeli, O. Onurand K. Uzer, Scoring and Predicting Risk Preferences,in *Behavior Computing: Modeling, Analysis, Mining and Decision,* Cao, L., Yu, P. S. (Eds), Springer(2012).

R. Agrawal, T. Imielinski and A.N. Swami, Mining association rules between sets of items in large databases,in *SIGMOD Conference*,P. Buneman and S.Jajodia, (Eds) (1993).

SAS. http://www.sas.com/

RapidMiner. https://rapid-i.com/content/view/181/190/

Weka. http://www.cs.waikato.ac.nz/ml/weka/

C. Borgeltand R. Kruse, Graphical models: methods for data analysis and mining, Wiley (2002).

E.N. Cinicioglu, G. Ertek, D. Demirerand H.E. Yoruk,A framework for automated association mining over multiple databases, in *Innovations in Intelligent Systems and Applications (INISTA), International Symposium, IEEE*, (2011).

A. Savasere, E. Omiecinski and S. Navathe, Mining for strong negative associations in a large database of customer transactions, in *Data Engineering, Proceedings., 14th International Conference, IEEE* (1998).

P.N. Tan, V. Kumar and H.Kuno, in *Western Users of SAS Software Conference* (2001).

I. Herman, G. Melanconand M.S. Marshall, Graph visualization and navigation in information visualization: A survey, *Visualization and Comp. Graphics*, 6 (2000)

M. Van Kreveld and B. Speckmann, Graph Drawing,Lecture Notes in Computer Science (LNCS), 7034 (2012).

R. Spence, *Information Visualization*, ACM Press (2001).

H. Ltifi, B. Ayed, A.M. Alimiand S. Lepreux,Survey of information visualization techniques for exploitation in KDD, in *Int. Conf. Comp. Sys.and App.*(2009).

C. Chen, Information Visualization, Wiley Interdisciplinary Reviews: Computational Statistics, 2 (2010).

[18] G. Ertek, A. Demiriz, and F. Çakmak, Linking Behavioral Patterns to Personal Attributesthrough Data Re-Mining, in <italic>Behavior Computing: Modeling, Analysis, Mining and Decision,<italic finished> Cao, L., Yu, P.S.(Eds.), Springer, (2012).

W.W. Cooper, L.M. Seiford and K. Tone, Introduction to Data Envelopment Analysis and Its Uses: With DEA Solver Software and References,Springer (2006).

S. Gattoufi, M. Oral and A. Reisman, Data envelopment analysis literature: A bibliography update (1951--2001), *Journal of Socio-Econ. Planning Sci.*, 38, pp. 159-229. (2004).

L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web,*The PageRank Citation Ranking: Bringing Order to the Web*(1999).

D.J. Watts and S. Strogatz, Collective dynamics of 'small-world' networks, *Nature*, 393 (1998).

T. Opsahl, F. Agneessensand J. Skvoretz, Node centrality in weighted networks: Generalizing degree and shortest paths,*Social Networks*, 32, pp. 245. (2010).

C. Christensen and A. Réka, Using graph concepts to understand the organizatıon of complex systems,*International Journal of Bifurcation and Chaos*, 17, pp. 2201-2214. (2007).

C.F. Chienand L.F. Chen,Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry, *Expert Systems with Applications*, 34, pp. 280-290.(2008).

J.R. Quinlan,Induction of decision trees, *Machine Learning*, 1(1), pp. 81-106.(1986).

Orange. http://orange.biolab.si/.

E.Alpaydin, *Introduction to Machine Learning*,The MIT Press(2010).

P.N. Tan, M. Steinbach and V. Kumar,*Introduction to Data Mining*, Pearson Addison Wesley (2006).

B. Bollobas,*Modern Graph Theory*, Springer (1998).

J.L. Gross and J. Yellen, Handbook of Graph Theory, CRC Press (2003).

M.E.J. Newman, *Networks: An Introduction*, Oxford University Press (2010).

WolframAlpha. http://www.wolframalpha.com/.

T.M.J. Fruchterman and E.M. Reingold, Graph Drawing by Force-Directed Placement, *Software – Practice & Experience*, 21(11), pp. 1129–1164. (1991).

E.M.Bonsignore, C. Dunne, D.Rotman, M. Smith, T. Capone, D.L. Hansen andB. Shneiderman, First Steps to NetViz Nirvana: Evaluating Social Network Analysis with NodeXL,in*International Symposium on Social Intelligence and Networking* (2009).

NodeXL. http://nodexl.codeplex.com/.

A.E. Akcay, G. Ertek and G. Buyukozkan, Analyzing the solutions of DEA through information visualization and data mining techniques: SmartDEA framework, *Expert Systems with Applications***39**, pp. 7763-7775, (2012).

T. Curk, J.Demsar, Q.Xu, G.Leban, U.Petrovic, I. Bratko, G. Shaulsky and B. Zupan,Microarray data mining with visual programming,*Bioinformatics*, 21(3), pp. 396-398. (2005).

R.D. Banker, A. Charnesand W.W. Cooper, Some models for estimating technical and scale inefficiencies in data envelopment analysis,*Management Science*. 30(9), pp. 1078–1092. (1984).