Luleå University of Technology (LTU)

Department of Computer Science, Electrical and Space Engineering

Digital Services and Systems

COURSE: D0025E

Data Mining Assignment Number 2: ASSOCIATION RULES

GROUP NO.: 2, MEMBERS: Ebbe BÄCKLINDER, Sophia DONATO, Simon JOHANSSON, Brandon LIU, Patrizio MANDELLI

LP1-2025

Abstract

This work aims to analyze relationships between items in a shopping basket in order to discover patterns, by using association rules. The dataset[1] contained almost 10,000 baskets that we got from Kaggle, and it was imported into Altair Rapidminer and Knime. No columns needed to be removed during preprocessing (except the one containing the total number of items per basket) as no basket were empty in the dataset; however, there were missing values for some attributes but it was logical since not every basket contained the same items. Using Altair Rapidminer and Knime, frequent itemsets and association rules were generated based on support, lift, and confidence measures. This work demonstrates how association rule mining can be applied for stores to uncover valuable insights on consumer behavior and provides practical guide to use data mining tools for Market Basket Analysis.

Highlights

This section explains shortly the key points of the assignment.

- Theory: The CRISP-DM process was carefully followed for this assignment. In Figure 1 shows the steps to follow the CRISP-DM process. Market Basket Analysis was applied to identify frequent itemsets and generate association rules.

Figure 1: CRISP-DM Process

- Practice: The key practical takeaways are preprocessing the dataset, and building working processes in Altair Rapidminer and Knime using correct nodes.

- Others: The key points include ensuring sufficient data for reliable results and reflecting on the discovered itemsets and association rules.

[1] https://drive.google.com/file/d/1RxIxB-ty5-tIHmsy3W3zJRdl_7ouLcni/view?usp=sharing

Contents

1 Phase I: Business Understanding

The purpose of this assignment is to analyze a dataset of grocery transactions in order to discover patterns of items that are frequently purchased together. In retail, this type of analysis is often referred to as market basket analysis and is widely used to support business decisions. By identifying which products are commonly associated, supermarkets can improve product placement, design more effective promotions,, and develop cross-selling strategies that align with real customer behavior and habits.

The business motivation is therefore to gain a deeper understanding of customer buying habits. Although the data set itself is transactional, the insights derived from association rules can inform practical actions such as optimizing shelf layouts, recommending complementary products, and managing inventory more efficiently. In this project, the specific data mining goal is to generate association rules that highlight significant product relationships, using thresholds for support, confidence, and lift to ensure that the results are both statistically valid and relevant for business understanding and use.

2 Phase II: Data Understanding

The dataset we used for this assignment consists of 9835 grocery transactions saved in a CSV file which we downloaded from kaggle. Each row represents a single shopping basket and contains the number of items purchased followed by up to 32 item entries. The items are typical products found in a supermarket, such as whole milk, yoghurt, butter, tropical fruit and bottled water.

The data is categorical in nature, as it records only the names of products without numerical details such as prices or quantities (it does contain one column documenting the amount of items purchased by that row in a numerical value but it is removed during the preprocessing). Because baskets vary in size, many of the item columns contain empty cells. These reflect the fact that some customers buy only a few items, while others buy more (max is 32). In both RapidMiner and KNIME, the data set appears with the same structure, and it is clear that the data is well suited to association rule mining.

A brief exploration shows that certain items occur much more frequently than others. Whole milk, for example, appears in more than 1/4th of all transactions, while products such as yogurt, roll and buns, and soda are also common. These observations confirm that the dataset contains strong and repeated patterns, making it appropriate for the subsequent search for association rules. Before any other step, it is vital to understand what structure the original file has and how to read in in RapidMiner and KNIME in such a way that it can be further analyzed with association processes.

3 Phase III: Data Preprocessing

The dataset we used for this assignment was consistent and clear, and more or less suited for the tools we used, requiring almost no preprocessing phase.

A first small step was required in both KNIME and RapidMiner, and consisted in removing the first column of the file, which contains the number of items per basket, which is not useful for the purpose of creating association rules, and which could even make the following steps fail in both softwares.

For RapidMiner, the preprocessing phase stopped at this moment since the follwoing tools allowed using a dataset with baskets as objects and each item in a different attribute/column, which exactly the way the CSV is constructed.

For KNIME, the following tools in the workflow allowing to create the association rules needed as input a datatable with transactions as collections, meaning that a transaction has the form of a single column containing a list of all included items. Data aggregation was thus done in

KNIME (see section 7) to achieve this correct representation.

In both softwares, other options were available concerning the shape of the input given to the association rules builder blocks, such as BitVectors in KNIME which is a special sequence of binary values where each bit represent the presence (1) or the absence (2) of a particular feature (here, item). To achieve this data shape, discretization and binarization would have been necessary, which a more complex task than data aggregation. We thus choose the simplest option.

4 Phase IV: Modeling

An association rule of the type $A \to B$, where A and B are itemsets of any dimension from a dataset Ω, provides information about the relationship between A and B. It can be analysed and understood using 3 main metrics :

Support

Support measures the frequency of the association rule:

$$S(A \to B) = \frac{N(A \cup B)}{N(\Omega)}, \quad S(A) = \frac{N(A)}{N(\Omega)}.$$

where $N(X)$ denotes the number of transactions in Ω containing itemset X.

- High support: the association rule is statistically significant.
- Low support: the association rule may be just a coincidence.

Confidence

Confidence measures the reliability of the association rule:

$$C(A \to B) = \frac{S(A \to B)}{S(A)}.$$

- High confidence: if A occurs, B is very likely to occur.
- Low confidence: if A occurs, B occurs only a few times.

Lift

Lift measures how much A occurs with B rather than with other items:

$$L(A \to B) = \frac{C(A \to B)}{S(B)}.$$

- $L > 1$: positive association.
- $L = 1$: no association.
- $L < 1$: negative association.

Steps for Discovering Association Rules

An association rule of the type $A \to B$ is discovered in two main steps:

1. **Frequent itemset discovery:** Based on a minimum support threshold, we find itemsets that occur frequently in the dataset.
2. **Association rule generation:** For each frequent itemset, we calculate the confidence and the lift. Minimum requirements for confidence and lift can also be specified.

5 Phase V: Evaluation

The evaluation phase assessed whether the generated association rules were statistically reliable and useful for decision making. Rules were reviewed based on support, confidence and lift to ensure they represented meaningful relationships rather than random patterns. Different support and confidence thresholds were tested, and the values of 0.02 and 0.3 provided a balanced set of interpretable rules.

6 Phase VI: Deployment

The results that we got can be used to understand how to optimize store layouts. For example the rules involving whole milk, root vegetables and other vegetables suggest that these items are often purchased together. To improve the store layout it would be a good idea to position whole milk and root vegetables closer to the fresh produce section to make it easier for customers to find all complimentary items in one trip. This can increase the likelihood of impulse purchases.

7 Tools Insights

7.1 RapidMiner Insights

Figure 2: RapidMiner workflow

- Read CSV: imports the Dataset.
- Select attribute : removes the column with the number of items per basket.
- FP-Growth : calculates the frequently occurring items in a transaction database (in this case, a transaction is the content of the shopping basket). It takes as entry a table where a transaction can have different form, as said earlier, we used the option 'Items in different columns' and just feed the CSV dataset to the operator. Another parameter to take into account is the 'min requirement' which can be either Frequency or Support, we used the support with different values. The output is a table of itemsets with the highest supports in the dataset, using the threshold defined earlier.
- Create Association Rule : creates association rules using as input a set of frequent (here in term of support) itemsets. One parameter to tune is the Confidence, for which we tested several values. The output are the discovered association rules (relatively to the threshold used).

7.1.1 Ease of Use :

RapidMiner is quite easy to use, once you understand correctly all the operators. The main challenge for this assignment was to understand correctly the FP-Growth and Create Association Rule operators, in particular the different options available, to prepare the data correctly and output useful results.

7.1.2 Efficiency :

The main advantage is the efficiency of the FP-Growth operator for mining large transaction dataset. The workflow can be reused and the parameters can be tuned without having to rebuild the whole workflow.

7.1.3 Costs:

The free version of RapidMiner allows to build a efficient and working process, as long as the dataset doesn't contain too many objects.

7.2 Knime Insights

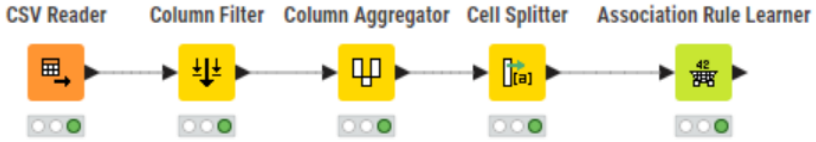

Figure 3: Knime workflow

- CSV Reader: imports the dataset.
- Column Filter : removes the column with the number of items per basket.
- Column Aggregator : aggregates all columns containing the different items in a single string column containing all items, separated by a comma. The main option was the aggregation method, for which we choose the concatenation.
- Cell Splitter : splits the string previously created into its elements (here, items) and creates a Collection (Set) out of them.
- Association Rule Learner : searches for frequent itemsets (with a threshold based on the Support value) and optionally creates association rules from them (if so, a threshold for the Confidence must be chosen), meaning that it does the work of FP-Growth and Create Association Rule at the same time.

7.2.1 Ease of Use:

KNIME is less straightforward than RapidMiner, however its use stays relatively easy. Once again, the main challenge was to understand the inputs/options required by the different operators in order to prepare the data correctly and obtain meaningful results.

7.2.2 Efficiency :

The Association Rule Leanrer operator combines the work of two operators in RapidMiner, making it more efficient. Once again, workflow can be reused and options can be modified without changing the whole workflow.

7.2.3 Costs :

The process can be entirely done using the free version of KNIME.

8 Answers to Assignment Questions

1. **What rules did you find?**

   The association rules show that when customers buy certain products, they are more likely to also buy other items. For example:

   - Tropical fruit, pork, butter, eggs, or sour cream often lead to buying other vegetables.
   - Root vegetables, yogurt, curd, or margarine often lead to buying whole milk.
   - Combinations like whole milk + root vegetables strongly point to other vegetables.

2. **What attributes are most strongly associated with one another?**

   Whole milk and other vegetables are the main items of association, connected to many other items.

   Looking at the lift, the strongest links are:

   - Whole milk + root vegetables → other vegetables (lift = 2.45)
   - Other vegetables + yogurt → whole milk (lift = 2.07)
   - Whipped/sour cream → other vegetables (lift = 2.08)
   - Pork → other vegetables (lift = 1.94)

3. **Are there products that are frequently connected that surprise you? Why do you think this might be?**

   Yes, some links are a bit surprising. For example, pork → other vegetables or sour cream → other vegetables. At first, meat or dairy don't seem directly related to vegetables. But this may happen because people often buy these together to make meals, like cooking meat with vegetables, or using sour cream in recipes that also need vegetables. So the connection makes sense once we think of how people cook.

4. **How much did you have to test different support and confidence values before you found some association rules?**

   We had to try a few different values before finding good rules:

   - If support was too high, we only got very common items like whole milk.
   - If support was too low, we got too many weak rules.
   - The same with confidence; too strict gave us almost nothing, too loose gave us rules that were not useful.

   So, it took a few tests to find a balance where the rules made sense. The final values chosen for support and confidence are respectively 0.02 and 0.3.

5. **Were any of your association rules good enough that you would base decisions on them? Why or why not?**

   Yes, some rules are strong enough to guide real store decisions. For example:

   - Whole milk + root vegetables → other vegetables (very strong rule)
   - Other vegetables + yogurt → whole milk

   A grocery store can use this to organize shelves in a smarter way:

   - Place root vegetables close to other vegetables and also near whole milk, since these are often bought together.
   - Put yogurt and vegetables closer to the dairy section, because shoppers who pick yogurt often also go for milk.
   - Items like pork, butter, and eggs could be positioned near vegetables, since the rules show they are connected in shopping baskets.

   By arranging shelves this way, the store makes it easier for customers to find items that naturally go together in meals. This can increase cross-selling (customers buying more items) and improve the shopping experience.

9 Conclusion

In this assignment, we have analyzed and explored the relationships between items in a shopping basket dataset. The dataset did not need to remove any columns since there were no empty baskets for preprocessing, with the goal of discovering meaningful association rules that can help stores make better business decisions. We applied association rule mining techniques in both RapidMiner and KNIME to identify frequent itemsets and generate rules based on support, confidence, and lift. The strongest associations we discovered were whole milk and other vegetables

that are frequently bought together with other products. Such as whole milk + root vegetables → other vegetables (lift = 2.45), or other vegetables + yogurt → whole milk (lift = 2.07). We got some other associations like other vegetables to pork(lift = 1.94), or other vegetables to sour cream(lift=2.08) that were less obvious at first but can be explained by common cooking habits. Overall, we learned that association rule mining is a valuable tool for identifying hidden buying patterns. These insights can be directly applied to store layout optimization and cross-selling strategies. For example, placing yogurt closer to milk and vegetables, or placing vegetables near whole milk. That could encourage customers to buy more items together by passing them on. We recommend that stores use these methods to improve product placement, marketing strategies, and sell more.