Association Rules Validation

Nadav Feige (205870421), Shimon Avraham (204623094)

Submitted as final project report for Tabular Data Science, BIU, 2023

Abstract

Association rule mining is a technique used to find relationships between variables in a dataset and present them as rules. These rules are often thought to be useful in making predictions about future events. However, commonly used measures such as support and confidence don't directly assess the predictive power of these rules. One challenge in association rule mining is to determine if the rules we find are actually meaningful and not just random, especially when the support values are very small. In this paper, the authors propose two methods for evaluating the predictive power of these rules. the first method involve dividing the data into training and testing sets, while the second compares the rules generated by the "Apriori" algorithm with those generated by a collaborative filtering recommendation model using "k-Nearest Neighbors".

Definitions

Itemset

A collection of zero or more items is termed an itemset. If an itemset contains k items, it is called a k-itemset (Hamida and Mohsen, 2015).

Association Rule

Let I = $\{i_1, i_2, i_3, ..., i_n\}$ be the set of all items and A,B be sets of items.
An association rule is an implication of the form $A \rightarrow B$, where:
- $A \subset I$, $B \subset I$

- A $\neq \emptyset$, B $\neq \emptyset$
- A $\cap$ B $= \emptyset$

Support

Given a transaction set D and an association rule r = A $\rightarrow$ B. the support of rule r is defined by:

$$\text{support(r)} := \text{P}(A \cup B) = \frac{\sum_{T \in D}[I_{A \cup B \in T}]}{|D|}$$

Confidence

$$\text{confidence(r)} := \text{P(B} \mid \text{A)} = \frac{P(A \wedge B)}{P(A)} = \frac{support(A \cup B)}{support(A)}$$

Problem Description

Association rule mining is a data mining technique that identifies correlations or patterns among items in a data-set. The technique involves two steps, the first of which involves identifying frequent item-sets whose occurrences exceed a predetermined threshold. In the next step, association rules are generated from the frequent item-sets based on the minimal confidence constraints. However, the extraction of association rules is faced with several challenges, including redundancy, a large number of rules, and relevance levels. Objective measures and subjective measures are often used to evaluate the interestingness of association patterns, but it is difficult for users to determine the appropriate threshold value. Additionally, the diversity of the data may result in small support values, which make it challenging to validate the credibility of the obtained rules. Therefore, different methods are required to validate the rules obtained from different algorithms and data-sets, such as filtering, testing, or applying specific criteria like coverage, leverage, lift, or strength.

Solution overview

Dataset:

We test our approach on 2 different datasets. Groceries and movies.

The Groceries Dataset

The dataset has 38,765 rows of the purchase orders of people from the grocery stores.

The Movies Dataset

The metadata for movies provides details on 45,000 films.

Description of the methods

The purpose of this step was to train the model and evaluate the strength of the rules, distinguishing them from random patterns. We began by running the Apriori algorithm on the entire dataset, followed by a division. Finally, we compared the test results with the Apriori algorithm's outcomes obtained from the undivided dataset. The third method compares the rules generated from the 'Apriori' algorithm with the rules generated by a collaborative filtering recommendation model using 'k-Nearest Neighbors' (kNN). If a rule is strong enough, it should be present in both sets of recommendations. This comparison allows identifying the strongest rules. In all cases, the goal is to validate the rules by ensuring that they are not biased or arbitrary and are accurate in predicting outcomes.

Apriori

An Apriori-based recommendation engine is a type of recommendation system that uses the Apriori algorithm to generate recommendations for users. The Apriori algorithm is a data mining technique used to discover associations between items in a dataset. It works by identifying frequent itemsets in a dataset, which are sets of items that occur together in transactions with a frequency above a certain threshold.

In the context of recommendation systems, the Apriori algorithm can be used to identify frequent itemsets of items that have been previously purchased or viewed by a user. Based on these frequent itemsets, the recommendation engine can then suggest other items that are frequently purchased or viewed together with the items in the user's history.

For example, suppose a user has previously purchased items A and B. The Apriori algorithm can be used to identify other items that are frequently purchased together with A and B, such as item C. Based on this association, the recommendation engine can then suggest item C to the user as a potential item of interest.

Overall, the Apriori-based recommendation engine is a powerful tool for generating personalized recommendations based on user behavior and item associations. However, it requires a large amount of data and computing power to be effective, and may not be suitable for smaller datasets or less powerful computing environments.

KNN

KNN (K-Nearest Neighbors) is a machine learning algorithm that can also be used as a recommendation engine. The basic idea behind the KNN recommendation engine is to recommend items to a user based on the preferences of other similar users.

In the KNN recommendation engine, each user is represented by a vector of ratings they have given to a set of items. The algorithm then computes the similarity between the target user and all other users in the dataset based on their rating vectors. The similarity metric can be any distance metric, such as Euclidean distance, cosine similarity, or Pearson correlation.

Once the similarities are computed, the algorithm selects the k most similar users to the target user, where k is a user-defined parameter. The algorithm then aggregates the ratings of the k most similar users for each item that the target user has not yet rated. Finally, the algorithm recommends the top-rated items to the target user.

One important aspect of KNN recommendation engines is that they require a dense and complete dataset. That is, each user should have rated a significant number of items, and there should not be many missing values in the dataset. Moreover, the computation of similarities can be computationally expensive for large datasets, so there are various optimizations, such as using approximate nearest neighbors algorithms, that can be used to speed up the computations.

In summary, the KNN recommendation engine is a simple and effective way to make personalized recommendations based on the preferences of

similar users.

Experimental evaluation

Division of the data for training and test

We performed the test on 'Groceries dataset' and checked whether the resulting rules were strong enough. We chose for the support values lower than one and presented in the table in the notebook additional indicators such as conviction for example. Finally, we presented the obtained differences in graphs.

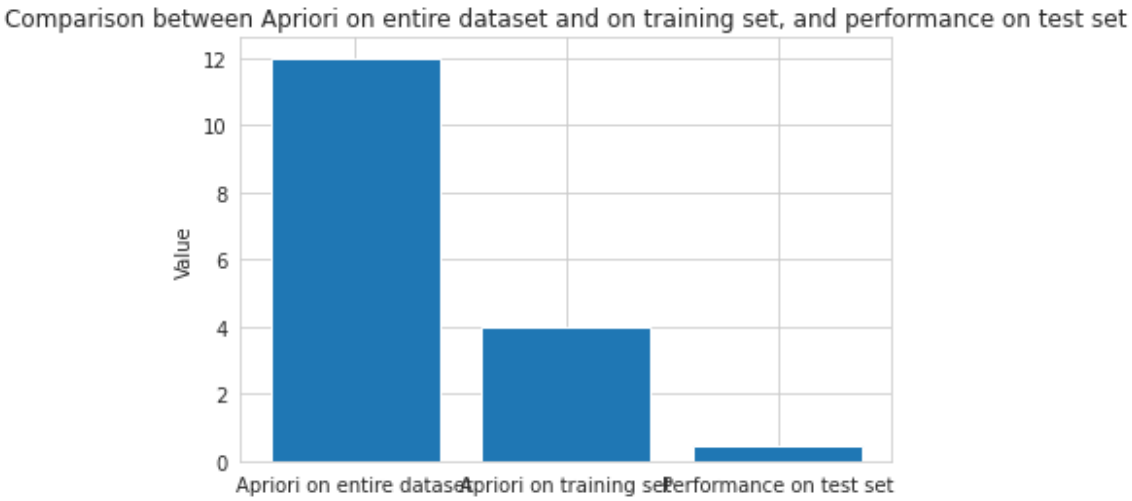

Figure 1: Caption

Comparing recommendations of kNN and 'Apriori' algorithms

We conducted a test on a collection of movies to identify the recommended movies for individuals who enjoyed a particular film. This involved combining two datasets - the movies dataset and the ratings dataset - using a shared ID. After merging the datasets, we selected a specific movie, 'Batman Returns,' and presented the top 5 recommended movies based on this selection.

Related work

The methods proposed in this paper are influenced by previous research on association rule mining. One of the methods involve dividing the data into training and testing sets, which are similar to the methods presented in previous articles. One of the articles, "Association rule discovery with the train and test approach for heart disease prediction" by Ordonez (2006) [2], introduced an algorithm that uses search constraints to reduce the number of rules and validates them on an independent test set. Another article, "Discovery of Temporal Association Rules Using Train and Test Approach" by Rajeswari et al. (2011) [3], proposed a similar method of pruning generated rules with validation in the testing phase. Both articles suggest advanced methods for further research, such as validation by cross-validation. However, this paper does not implement this method. The Second method proposed in this paper involves comparing rules generated by the "Apriori" algorithm with those generated by a collaborative filtering recommendation model using "k-Nearest Neighbors" (kNN). This method is based on the proposal in "A Novel and Efficient KNN using Modified Apriori Algorithm" by Agarwal et al. (2012) [1], which suggests modifying the Apriori algorithm to classify data for K-nearest neighbor. The presented method is inspired by this article.

Conclusion

In conclusion, we proposed two methods for evaluating the predictive power of association rules: dividing the data into training and testing sets, and comparing the rules generated by the Apriori algorithm with those generated by a collaborative filtering recommendation model using k-Nearest Neighbors. We applied these methods to two datasets, Groceries and Movies, and demonstrated the effectiveness of the approaches in identifying meaningful rules. We highlighted the challenges associated with validating association rules, particularly when support values are small, and suggests that methods such as filtering, testing, or applying specific criteria like coverage, leverage, lift, or strength may be useful in validating the credibility of the obtained rules. These methods can help ensure that the rules are not biased or arbitrary and are accurate in predicting outcomes.

References

[1] Ritika Agarwal, Barjesh Kochar, and Deepesh Srivastava. A novel and efficient knn using modified apriori algorithm. *International Journal of Scientific and Research Publications*, 2(5):2250–3153, 2012.

[2] Carlos Ordonez. Association rule discovery with the train and test approach for heart disease prediction. *IEEE transactions on information technology in biomedicine*, 10(2):334–343, 2006.

[3] AM Rajeswari, S Nithya Shalini, and C Deisy. Discovery of temporal association rules using train and test approach.