Market Basket Analysis

Sam Caekaert

09-21-2020

Introduction

This document provides an overview of the github repo located at `https://github.com/samcaek/market-basket-analysis`. The corresponding SSH clone is:

```
$ git clone git@github.com:samcaek/market-basket-analysis.git
```

Frequent Itemsets

The github repo calculates the frequent item sets associated with a set of market data. The frequent itemsets are generated using the apriori algorithm from (Zaki and Meira Jr [2019]).

Association Rules

The project also generates a set of association rules corresponding to a minimum confidence level using algorithm 8.6 from Zaki and Meira Jr [2019]. The association rules are then ranked based on the cumulative rank of three measures. The first measure is the confidence, defined as: $\text{conf}(X, Y) = \frac{\sup(X \cup Y)}{\sup(X)}$. The second measure is lift: $\text{lift}(X, Y) = \frac{\text{conf}(x,y)}{\sup(y)}$. The final measure is the number of items in the transaction: $\text{size}(X, Y) = |X| + |Y|$. The ranking system is used to normalize the measures which allows them to be combined in a way that is somewhat equal.

How to run

The code can be ran using any dataset that contains a single item per row. With this format, the transaction column will not be unique. The script currently transforms this data into the binary matrix from Zaki and Meira Jr [2019]. Two examples of this data can be seen in the files "book_example.csv" and "txn_by_dept.csv".

To execute the code using default settings, simply clone the repository and run:

```
$ python3 market-basket-analysis.py
```

By default, the code runs the example presented in Chapter 8 of Zaki and Meira Jr [2019]. This is simple example with a minimum support of 3 and minimum rule confidence of 0.5.

To use other data, the script allows for various paramaters to be passed from the command line. These include –dataset, –transaction-column,–item-column, –minsup, –minconf, –num-rules, –max-rows. The following is an example call using these paramaters:

```
$ python3 market-basket-analysis.py --dataset "txn_by_dept.csv"
↪  --transaction-column 0, --item-column 1, --minsup 4,
↪  --minconf 0.4, --num-rules 5, --max-rows 200
```

The above example returns the following 5 highest ranked rules:

1. ['0973:CANDY', '0982:SPIRITS'] —> ['0983:WINE', '0991:TOBACCO']
2. ['0982:SPIRITS', '0983:WINE', '0991:TOBACCO'] —> ['0973:CANDY']
3. ['0973:CANDY', '0982:SPIRITS', '0991:TOBACCO'] —> ['0983:WINE']
4. ['0983:WINE', '0991:TOBACCO'] —> ['0973:CANDY', '0982:SPIRITS']
5. ['0973:CANDY', '0982:SPIRITS', '0983:WINE'] —> ['0991:TOBACCO']

References

M. J. Zaki and W. Meira Jr. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge University Press, 2019.