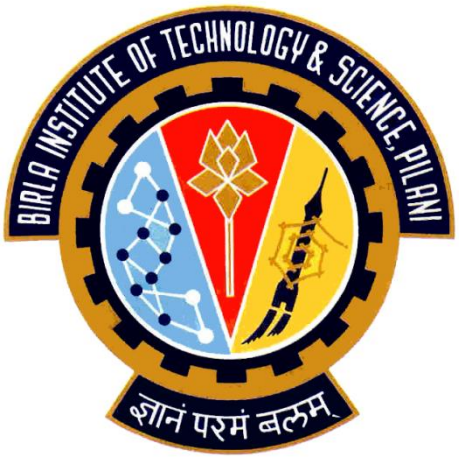
Birla Institute of Technology and Science-Pilani, Hyderabad Campus

First Semester 2017-18



Data Mining (CS F415)

Association Rule Mining

by

Shubham Jha	2015A3PS0288H
Praneet Mehta	2015A3PS0342H
Abhinav Jain	2015A7PS0174H

Under the guidance of

Mrs. Aruna Malapati

Dataset

Groceries Market Basket Dataset
http://www.sci.csueastbay.edu/~esuess/classes/Statistics_6620/Presentations/ml13/groceries.csv

Number of transactions: 9835

Number of unique items: 169

Pre-processing done on data

Groceries.csv file was read transaction by transaction and each transaction was saved as a list. A mapping was created from the unique items in the dataset to integers so that each item corresponded to a unique integer. The entire data was mapped to integers to reduce the storage and computational requirement. A reverse mapping was created from the integers to the items, so that the item names could be written in the final output file.

Formulas Used

Confidence (X -> Y) = support (X U Y) / support (X)

Support (X, Y) = support-count (X, Y) / total dataset size

We have used support instead of support count because computations with integers are faster than that of floating point numbers.

Support (X) = Support count (X) / Total number of transactions

Results for different for values of support and confidence

Confidence/Support	No. of frequent itemsets	No of rules
High confidence(MIN_CONF=0.5) High support count(MINSUP=60)	725	60
Low confidence(MIN_CONF=0.1) High support count(MINSUP=60)	725	1189
High confidence(MIN_CONF=0.5) Low support count(MINSUP=10)	11390	4187
Low confidence(MIN_CONF=0.1) Low support count(MINSUP=10)	11390	35196

frequent_itemset.txt and association_rules.txt for different MIN_CONF and MINSUP values can be found in the RESULTS folder

Observation

Most of the rules we generated have a common item (*whole milk* and *other vegetables*) on the consequent side. This happens when any item is very frequent in the transactions. This can be avoided by using ***lift*** instead of confidence.

Lift (X -> Y) = support (X U Y) / support (X) * support (Y)