RAMAIAH
Institute of Technology

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Data Mining and Machine Learning- CS

Submitted by

Amaresh TN	1MS19CS012
Anish S	1MS19CS018
Chinmay S Nadgir	1MS19CS038
Gagan Nischal A	1MS19CS044

Submitted as a part of the 20 mark component for the subject Data Mining and Machine Learning

Under the Supervision of

Vandana Sudhakar

Assistant Professor

RAMAIAH INSTITUTE OF TECHNOLOGY

(Autonomous Institute, Affiliated to VTU)

BANGALORE-560054

www.msrit.edu, March-July 2022

Ramaiah Institute of Technology

Abstract

Data is used constantly by companies present all over the world which have various innumerable sources through the internet. This available data is present in a raw format which can be made useful by preprocessing the data which can then be understood by the machines.

Data preprocessing is a technique in data mining and data analytics that considers the available raw data and transforms it into a format that is useful and efficient for the computers. We make use of data sets in order to train the machine learning models and it's quite common to hear the phrase “garbage in, garbage out” which means if we use data that is dirty then we end up having a model that is not trained properly.It's very important to note that clean preprocessed data is at times more important than the most algorithms itself.

There are many steps involved in data preprocessing like -

1.Data cleaning- it’s done to handle parts of the data that are not present or that are irrelevant. It has 2 subdivisions which are

i)Noisy data

ii)Missing data

2.Data transformation- it’s a process of converting data into a format that is suitable for the mining process. The different ways are -

i)Normalization

ii)Attribute Selection

iii)Discretization

iv)Concept Hierarchy Generation

3.Data Reduction- this is done in order to enhance the storage efficiency and reducing the analysis cost while handling data of huge volume.The different steps followed for data reduction:

i)Data Cube Aggregation

ii)Attribute Subset Selection

iii)Numerosity Reduction

iv)Dimensionality Reduction

In our current project we have used all the above steps in order to preprocess our data efficiently.

Now the algorithms that we have used in our project are:

1. Apriori algorithm
2. K-means algorithm

Apriori algorithm is the foundational algorithm in basket analysis. Basket analysis means the understanding of a customer's basket when they shop.Aim is to find out the various amalgamations of items that are often bought together, which is called Frequent Itemset Mining.We use the Apriori algorithm in order to preprocess our grocery shopping dataset in this project.

The topic Clustering is a very frequently used technique in order to analyze data and get a sense of clairvoyance on the data's structure. Kmeans behaves as an iterative algorithm which attempts to segregate the available dataset into KPre-defined unique subgroups that are non-overlapping (clusters) and ensures that each point of data relates to a particular group. We have used the K-Means algorithm in our project on the insurance policy dataset.

Dataset description

A Dataset can be defined as a grouping / collection of data that can be used for various algorithms. It's important to have the availability of superior-quality, congruous metadata which becomes important in searching, comprehensing, and reiterating the data that's scientific.

In our current project we have mainly used 2 datasets, they are - Grocery shopping dataset and Insurance policy dataset.

Grocery shopping dataset was obtained from kaggle which consisted of a dataset that had 38765 rows of customer purchasing orders collected through a wide number of grocery shops. These purchases can be used for scrutinizing and market basket analysis can also be performed using the association rules. In our project we can see that the groceries dataset does not have headers and thus we specify the parameter header='none' while converting it into a dataframe. We have used a matrix that has 9835 rows and 32 columns as our grocery dataset.

We have used **Insurance dataset** that is a collection of age, sex , body mass index(BMI), children, smoker, region and expenses. The insurance dataset was vital in order to perform and apply clustering which is why we have applied the K-means algorithm on this dataset. A conversion of 'SEX' and 'SMOKER' which is a binary categorical value to 0 and 1 has been done. At the same time converting the 'REGION' which is a multi class categorical variable to one hot encoded format and merging it back to the original data frame has also been done. The matrix of our dataset has 1338 rows and 7 columns where the division of the given dataset into 30% and 70% partition has been done for our use.

Algorithm Description

Apriori Algorithm

This is an algorithm that is used to calculate the associativity rules amidst various items. The Apriori algorithm makes use of mainly 2 steps which are, "join" and "prune" in order to decrease the search space. The algorithm was given such a name called Apriori as it utilizes the earlier knowledge of common itemset descriptives. Application of an iterative or level-wise search approach is done where k-frequent itemsets are utilized to search for k+1 itemsets. It acts as a tool for finding probability using association rule mining like for eg , it helps us in analyzing that customers who purchased item A may also most likely be interested in buying item B.The various steps of this Apriori algorithm are:

Step 1. Calculating the support for every single item

Step 2. Determining the support threshold

Step 3. Choosing the repetitive items

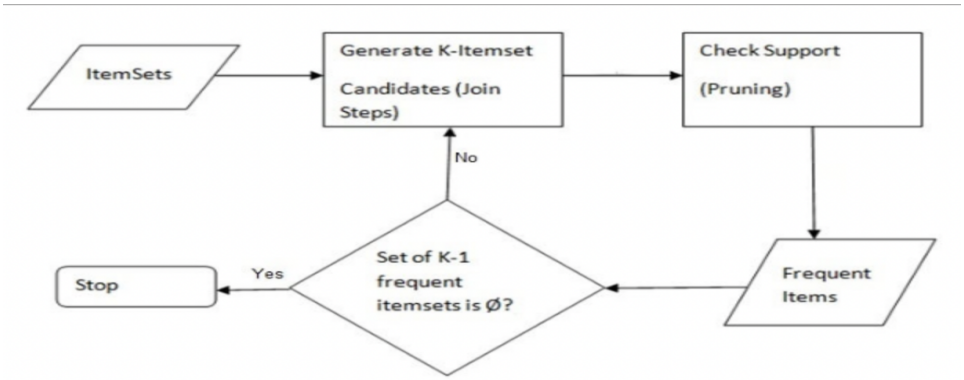
Step 4. Searching support of repetitive itemsets

Step 5. Repeat for bigger sets

Step 6. Produce Association Rules and calculate confidence

Step 7. Calculate lift using the formula below

$$lift = \frac{P(X \cap Y)}{P(X) * P(Y)}$$



K-Means Algorithm

This algorithm is a Clustering algorithm which is an Unsupervised Learning algorithm that helps in grouping unlabeled dataset into various clusters. Over here K is used to define the quantity of pre-defined clusters that have to be generated in available processneed , like for eg if K=5, there are going to be 5 clusters and when K=6 there are going to be 6 clusters and so on.

This clustering algorithm helps in calculating centroids and goes on iterating until an optimal centroid is found. The various steps are :

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids.

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

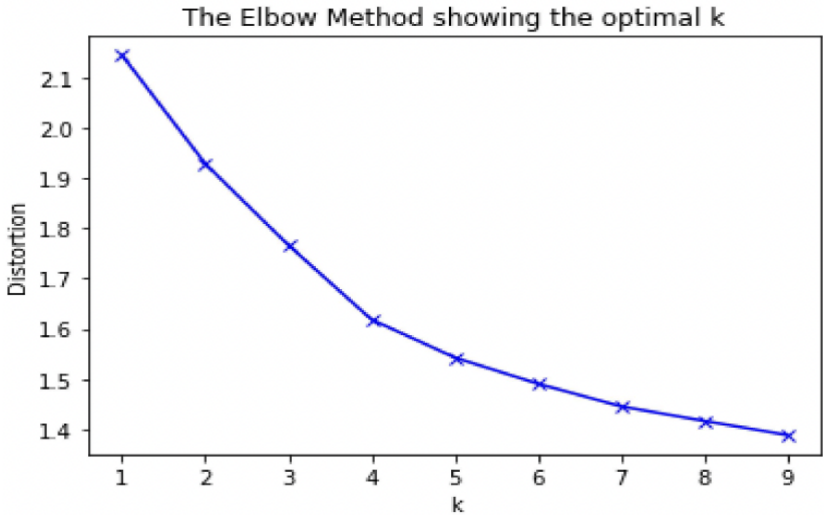
Step-4: Calculate the variance and place a new centroid of each cluster.

Step-5: Reassign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

Step-7: The model is ready.

In this a major component that is needed for all unsupervised algorithms is to find out the optimal quantity of clusters which has the data that can be clustered as per our needs. In order to do this, the **Elbow Method** acts as a very famous method that helps to find the optimal value of k.



Results and Inferences

Inferences of Apriori Algorithm:

```
Rule:whipped/sour cream-->berries
support:0.017
confidence: 0.3617021276595745
lift: 4.822695035460994
**********

Rule:root vegetables-->butter
support:0.017
confidence: 0.3617021276595745
lift: 3.2585777266628333
**********

Rule:citrus fruit-->sugar
support:0.014
confidence: 0.32558139534883723
lift: 3.3565092304003836
**********

Rule:dessert-->yogurt
support:0.013
confidence: 0.3939393939393939
lift: 3.077651515151515
**********
```

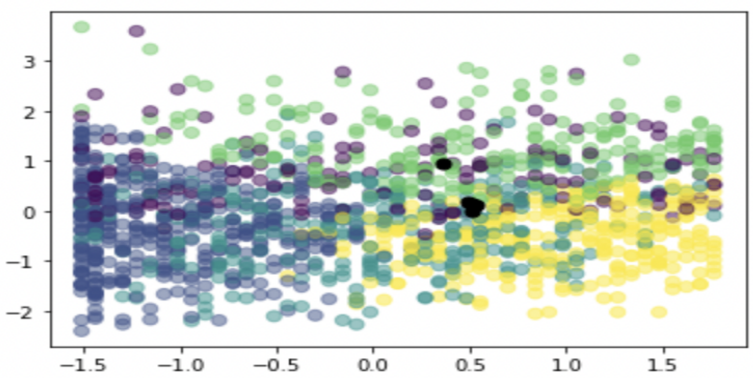
We see that the apriori algorithm works well to map the association rules so as to come up with the right patterns for predictions of the sales for the business owner . Thus he could infer from the results of the algorithm what group of items he can pair next to each other to increase his sales and make right business decisions to take his revenue towards the right direction.

Note : Confidence , Support and length can be varied as required. Nan is also used as an item in our dataset which means that the customer hasn't bought anything .

Inferences of K-means Algorithm:

Thus the K-means clustered data is used as a customer segmented result for the given insurance dataset and helps the company make decisions based on the cluster a given person belongs to and thus calculate the premiums analyze the risk ratios of that individual and thus come up with cost cutting decisions in certain cases and thus improve the revenue scale of the corporation.

2-D representation of clusters



We project our obtained clusters to 3 dimensions to analyze it more easily .

3-D representation of clusters