International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 3 Issue 5, August 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Delivery Feet Data using K-Mean Clustering with Applied SPSS

San San Nwe, Khin Khin Lay, Myint Myint Yee

Associate Professor, University of Computer Studies, Maubin, Myanmar

***How to cite this paper**:* San San Nwe | Khin Khin Lay | Myint Myint Yee "Delivery Feet Data using K-Mean Clustering with Applied SPSS" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-5, August 2019, pp.1944-1945, **https://doi.org/10.31142/ijtsrd26816**

IJTSRD26816

ABSTRACT

Data mining refers to extracting or mining knowledge from large amounts of data. Many people treat data mining as a synonym for another popularly used term, knowledge discover from data or KDD. Data can be mined such as relational databases, data warehouses, transactional databases, advanced data and information systems and advance applications. The construction of clustering model which classify with car driving analysis using K-mean clustering algorithm. The dataset was downloading from Google.com.

***KEYWORDS:** Data Mining, K_mean clustering algorithm, SPSS.*

1. INTRODUCTION

Unsupervised Learning used clustering methods. Clustering is the process of grouping a set of physical or abstract data objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in another. Cluster analysis is a popular data discretization method. A clustering algorithm can be applied to discretize a numerical attribute. There are categorization of major clustering methods are Partitioning methods, Hierarchical methods, Density based methods, Grid based methods, Model based methods, Clustering high dimensional data, constraint based clustering.

1.2 SPSS

SPSS standing for Statistical Package for the Social Sciences is powerful, user friendly software package for the manipulation and statistical analysis of data.

The package is particularly useful for students and researchers in psychology, sociology, psychiatry, and other behavioral sciences, containing as it does an extensive range of both univariate and multivariate procedures.[1]

1.3 SPSS clustering

The SPSS classify methods are Two step cluster, K-mean cluster, Hierarchical cluster and Cluster Silhouettes. This paper we used K-mean clustering algorithm, K-mean clustering algorithm defines the centroid of the cluster as the mean value of the points within the cluster. [3]

2. Algorithm

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\|x_i - v_j\|\right)^2$$

Where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

Algorithmic steps for k-means clustering

Let X = $\{x_1,x_2,x_3,\ldots\ldots..,x_n\}$ be the set of data points and V = $\{v_1,v_2,\ldots\ldots.,v_c\}$ be the set of centers.

1. Randomly select 'c' cluster centers.
2. Calculate the distance between each data point and cluster centers.
3. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers...
4. Recalculate the new cluster center using:

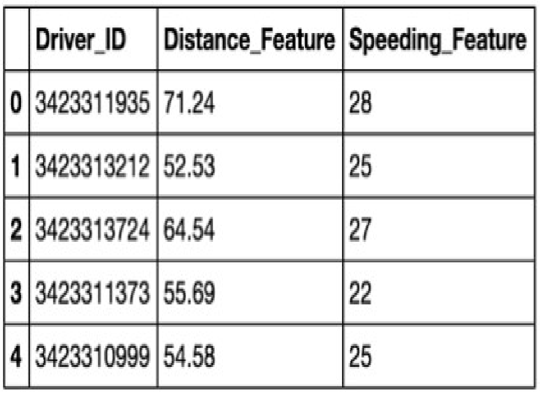
$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, 'ci' represents the number of data points in ith cluster.

5. Recalculate the distance between each data point and new obtained cluster centers.[2]

3. Testing

Sample dataset

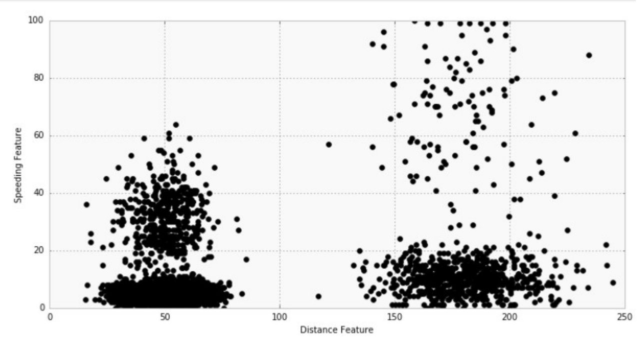


	Driver_ID	Distance_Feature	Speeding_Feature
0	3423311935	71.24	28
1	3423313212	52.53	25
2	3423313724	64.54	27
3	3423311373	55.69	22
4	3423310999	54.58	25

3.1 Choose K and Run the Algorithm

The chart below shows the dataset for 4,000 drivers, with the distance feature on the x-axis and speeding feature on the y-axis.

Start by choosing *K*=2 computations as shown below:



3.2 Review the Results

The chart below shows the results. Visually, you can see that the *K*-means algorithm splits the two groups based on the distance feature. Each cluster centroid is marked with a star.

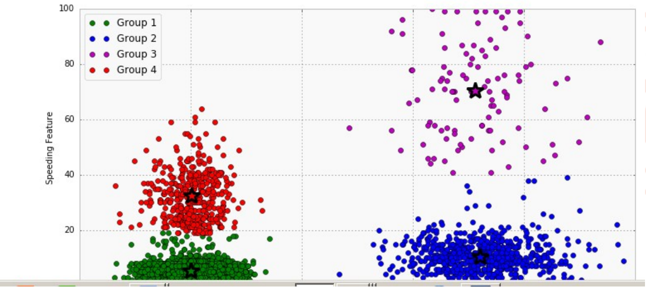
- Group 1 Centroid
  Distance feature input = 50miles and speeding feature = 50.2
- Group 2 Centroid
  Distance feature input = 180.3 miles ad speeding ffeature = 10.5

Using domain knowledge of the dataset, we can infer that Group 1 is urban drivers and Group 2 is rural drivers.

3.3 Iterate Over Several Values of K

Test how the results look for *K*=4.

The chart below shows the resulting clusters. We see that four distinct groups have been identified by the algorithm; now speeding drivers have been separated from those who follow speed limits, in addition to the rural vs. urban divide. The threshold for speeding is lower with the urban driver group than for the rural drivers, likely due to urban drivers spending more time in intersections and stop-and-go traffic.



3.4 Data Analysis View

The process of using domain knowledge to choose which data metrics to input as features into a machine learning algorithm. *K*-means clustering is using meaningful features that capture the variability of the data is essential for the algorithm.

Feature transformations, particularly to represent rates rather than measurements, can help to normalize the data. For example, in the paper how many total distance driven had been used rather than mean distance per day, then drivers would have been grouped by how long they had been driving for the company rather than rural vs. urban.

4. Conclusion

A Clustering method has been proposed in the unsupervised learning method. The deliver driving rate using with cluster algorithm using SPSS. SPSS is data analysis tools are valuable in social science. It is very good for presentation report by graph design. One possible outcome is that there are no organic clusters in the data; instead, all of the data fall along the continuous feature ranges within one single group. It may need to revisit the data features to see if different measurements need to be included or a feature transformation would better represent the variability in the data. In addition, you may want to impose categories or labels based on domain knowledge and modify your analysis approach.

References:

[1] A handbook of statistical analyses using SPSS/ Sabine, Landau, Brain S. Everitt, IBSN1-58488- 369-3[book style]

[2] Data Mining Concept and Techniques[Jiawei Han , Micheline Kamber , Jian Pei] [2] Student_user_guide_for_spss[BarnardCollege/ Department of Biological Science]

[3] IBM SPSS Statistics 23 part1[Information Technology Service, Winter 2016,version1]

[4] Handbook of Statistical Analysis & Data Mining Applications[Nisbet, Elder & Miner 2009-06-05]

[5] Doing Statistic with SPSS [Alistair W. Kerr , Howard K. Hall, Stephen A. Kozub]