Available online at www.sciencedirect.com

ScienceDirect

Procedia Economics and Finance 38 (2016) 70 – 76

Procedia Economics and Finance

www.elsevier.com/locate/procedia

Istanbul Conference of Economics and Finance, ICEF 2015, 22-23 October 2015, Istanbul, Turkey

An Application of Fuzzy Clustering on Prevalence of Youth Tobacco Survey

Hazel KAVILI[a]*, Gülhayat GÖLBAŞI ŞİMŞEK[b]

[a]*Hazel KAVILI, Yıldız Technical University, Istanbul, 34220, Turkey*
[b]*Gülhayat GÖLBAŞI ŞİMŞEK, Yıldız Technical University, Istanbul, 34220, Turkey*

Abstract

The foremost avoidable cause of disease and death is tobacco consume in almost every country and nearly all tobacco consume begins during youth and young adulthood. The use of tobacco products has increased in years and nearly half of tobacco consumers use more than two tobacco products. The paper is about the application of fuzzy clustering on the data of young people's attitude toward tobacco products. Fuzzy clustering was used to cluster people into a number of groups based on their use, tendency and intention. In our experiments, we used Fuzzy C-Means algorithms and it was calculated by using R Studio's packages. In our analysis, we have considered 7 questions from National Youth Tobacco Survey 2013 Questionnaire. Questions are about use of different tobacco products and thoughts about tobacco. We examined 267 subjects who were randomly selected. As a result of our analysis we came to the conclusion that most young people are not interested in smoking any kind of tobacco products. Most of them have not tried any kind of tobacco but some of them have tendency to smoke in years.

Peer-review under responsibility of the Organizing Committee of ICEF 2015.

Keywords: Fuzzy Clustering; Fuzzy Logic

* Hazel KAVILI Tel.: +90-212-383-4421;
E-mail address: hazzelkavili@gmail.com

Peer-review under responsibility of the Organizing Committee of ICEF 2015.
doi:10.1016/S2212-5671(16)30178-2

1. Introduction

Data mining and data analysis underlie many computing implementation, whether in a design phase or as part of their online process. The key element in different type of process (either exploratory or confirmatory) is grouping or classification of measurements based on goodness of fit to acknowledged model or natural groupings reveal through analysis.

Clustering is grouping a set of objects in a way that objects in the same group are more similar to each other than to those in other groups. Clustering algorithms are used in many study fields such as image analysis, machine learning, bioinformatics, text analysis etc. For cluster analysis, there is not one specific algorithm. It can solved by various algorithms that show significant differences in their notation of what constitutes a cluster and how efficiently find them. Besides, in some problems, data analyst has a little prior information about the data and analyst need to make a few assumptions about the data as possible.

The goal of this paper is to explain fuzzy clustering algorithm and the logic behind this. Also, making a different approach for National Youth Tobacco Survey (2013) by using fuzzy clustering algorithm.

1.1. Clustering Algorithms

Clustering algorithms can be classified ground on their cluster type such as hard clustering, soft clustering, centroid models, hierarchical clustering etc. The most convenient clustering algorithm for a certain problem should be determined experimentally or it can be chosen for mathematical reasons.

Clustering algorithms are procedures that make use of similarity or dissimilarity matrix to group objects homogeneously within cluster and heterogeneously between clusters. The most popular algorithms can be grouped as hierarchical or non-hierarchical clustering. Hierarchical clustering is connectivity based clustering. The idea of this method is the data objects being more related to nearby objects than the objects farther away. The connection between these objects can be explained by distance. At different distances, there will be different clusters, which can be shown using a dendogram. Hierarchical algorithms do not provide single partitioning of the data set, instead of that, they provide an extensive hierarchy of clusters that merge with each other at certain distances.

One of the most popular clustering algorithms is K-Means Clustering. The aim in the clustering algorithm is to separate the data example into small clusters or small subsets. In K-means algorithm, the data objects allowed to belong only one cluster. The algorithm places the statistically similar data objects into the same cluster. The center of the cluster represents the cluster. The letter “k” in the K-means algorithm, represents the number of clusters. The algorithm uses squared error function to minimize the error and search for the appropriate “k”. The given “n” number of data objects are placed to “k” clusters by minimizing squared error function. Therefore, the similarity of clusters measured by the proximity of the mean of the objects in the clusters. And this is the center of gravity of cluster. The object in the cluster is representative object of the cluster and it is called medoid. The algorithm has such steps to get clusters homogeneously within and heterogeneously between clusters:

1- Specify cluster centers
2- Classify objects by distances (dissimilarity or similarity)
3- Specify cluster centers after classification objects
4- Repeat 2^{nd} and 3^{rd} steps

In the topic of clustering, distance measure between data points is an important component. The most commonly used distance measures are Euclidean, Minkowski, Manhattan. Euclidean is commonly used for evaluation of the