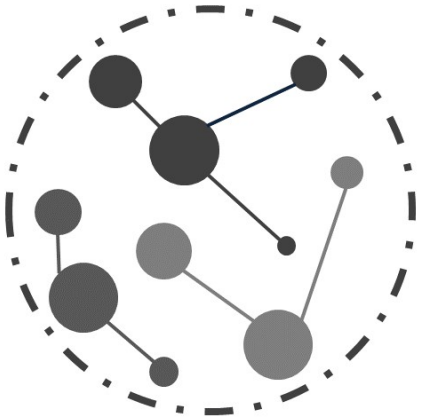
Clustering algorithms in the touristic sector:

Case study on travel agency data

Argyro Sioziou

A thesis presented for the degree of
Management Science and Technology



Management Science and Technology
Athens University of Economics and Business
Greece
20/02/2020

Clustering algorithms in the touristic sector

Case study on travel agency data

Argyro Sioziou

Abstract

Contents

Chapter 1

INTRODUCTION

1.1 Research Motivation

1.2 Research Methology

Chapter 2

BACKGROUND

2.1 Overview

2.1.1 Machine Learning

Machine learning is the field of study that focuses on training machines (e.g. computers) to identify patterns and derive logical conclusions (Bishop, 2006). It is technically an imitation of the human learning process and can be divided into two main different types, supervised and unsupervised learning.
Supervised learning is used to classify instances to already known categories based on their characteristics. Algorithms that belong to this type, are first trained on an already labeled with the possible categories dataset and afterward use their gained knowledge to classify new unlabeled datasets.
Unsupervised learning on the other side is used to group data without knowing the labels beforehand and without any training proceeding. This type of learning allows the model to learn by itself. Usually, a person with knowledge in the sector is needed to interpret and extract the knowledge from the created groups.

2.1.2 Clustering

Clustering or cluster analysis is a type of unsupervised machine learning. It groups instances to create coherent sets based on their similarities and dissimilarities. The aim is that the instances that belong to the same sets are as much alike and as much different from the instances of the other sets as possible (Dunham, 2002; Tan et al., 2005).

2.1.3 Types of Clustering

Clustering algorithms can be categorized based on two criterions. The first one is the relationships between the produced clusters of one iteration with the ones of a previous iteration. The second one is the number of clusters an instance can be a part of.
Based on the first criterion, Tan et al. (2005) describe the following types:
Partitional clustering is a type of clustering which allows no overlapping between two or more sets of clusters, hence partitions the initial set to independent sets. This means that each instance can be part of exactly one cluster.
Hierarchical clustering produces clusters in each iteration which are subsets of a cluster of the previous iteration. This can occur by starting from one cluster, which contains all instances, and repeatedly partition the available clusters to even smaller ones.
Tan et al. (2005) mention three more types of clustering in their book, which are formed based on the second criterion and are the following:

Exclusive clustering signifies that an instance can only be a part of only one cluster.
Overlapping or non-exclusive clustering signifies that an instance can be part of more than one cluster.
Fuzzy clustering signifies that every instance belongs to all the clusters (Tan et al., 2005).

2.2 K-means

K-means is a prototype-based, iterative algorithm in which instances are assigned to a cluster in each iteration. The basic algorithm requires as input K points, called centroids, where K is the number of the desired clusters. To define the cluster that one instance belongs to the algorithm calculates its distance from all the centroids and assigns it to the closest one. Finally, using the produced clusters calculates the new centroids and repeats until the desired set is reached. To calculate the distances and the new centroids the cluster mean needs to be defined and calculated (Dunham, 2002; Tan et al., 2005).
In the simple case where there is only one numerical value describing each instance the cluster mean can use the basic mean definition from statistics as follows:
–TO DO–

2.3 Applications in the touristic sector

2.3.1 Benefit Segmentation

As Dolnicar (2008) states companies all around the globe use market segmentation to target their audiences more effectively and consume their resources more efficiently. Cluster analysis is a useful tool for the creation of those segments to create groups of people that seek similar benefits from their travel experiences. These groups not only help in the formation of marketing strategies but also to products that offer higher fulfillment to their consumers (Dolnicar, 2008). This type of segmentation was introduced by Haley (1968) in 1968 and is called benefit segmentation. The need for the creation of a new way of grouping target groups arose from the fact that tourists' behavior is mainly determined by the benefits they seek to satisfy and not by descriptive factors (Haley, 1968, p. 31).
A sample application of benefit segmentation using cluster analysis describe Pesonen et al. (2011) in their article. Data about tourists were gathered in the region of Savonlinna, Finland during the period with the highest footfall of the year through electronic questionnaires. The research aims to find the different segments of tourists based on the benefits they seek and then examine the interest of each segment on wellness holidays. For the segment formulation, two algorithms were used, one hierarchical to find the best number of clusters, and K-means to create the clusters. The solution they proposed contained four distinguishable clusters. From the four clusters, two of them seemed to have a higher preference for wellness services, as designated by Pesonen et al. (2011) 'Culturals' and 'Sightseers'. Both clusters portrayed people that seem to favor attractions and also have a tendency to go back to the same destination for their vacation. Additionally, the 'Culturals' show a preference for cultural activities, where 'Sightseers' show a preference for sightseeing activities. The illustrated interpretation of the results is that the previous approach, which promoted wellbeing products based on nature, might not be the most appropriate. The suggested alternative claims that combining wellbeing with history, culture, diverse experiences and attractions could be more efficient (Pesonen et al., 2011, pp. 308–312).

Chapter 3

CASE STUDY: CLUSTER ANALYSIS ON TRAVEL AGENCY'S DATA

Bibliography

Bishop, C. M. (2006). *Pattern recognition and machine learning.*

Dolnicar, S. (2008). A review of data-driven market segmentation in tourism. *Journal of Travel & Tourism Marketing*, *12*(1), 1–22. https://doi.org/10.1300/J073v12n01_01

Dunham, M. H. (2002). *Data mining: Introductory and advanced topics.*

Haley, R. I. (1968). Benefit segmentation: A decision-oriented research tool. *Journal of Marketing*, *32*(3), https://doi.org/10.1177/002224296803200306, 30–35. https://doi.org/10.1177/002224296803200306

Pesonen, J., Laukkanen, T., & Komppula, R. (2011). Benefit segmentation of potential wellbeing tourists. *Journal of Vacation Marketing*, *17*(4), https://doi.org/10.1177/1356766711423322, 303–314. https://doi.org/10.1177/1356766711423322

Tan, P.-N., Kumar, V., & Steinbach, M. (2005). *Introduction to data mining.*