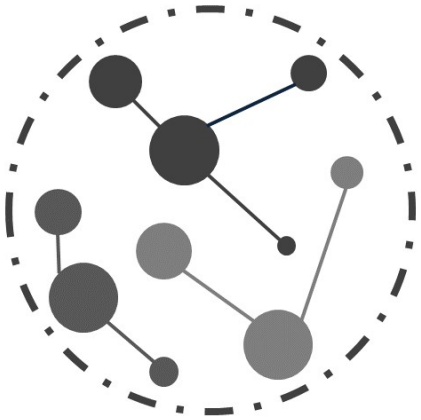
Clustering algorithms in the touristic sector:

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. It is technically an immitation of the human learning process and can be divided in two main different types, supervised and unsupervised learning. (Bishop, 2006)
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 afterwards 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 on 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 in order to create coherent sets based on their similiraties and disimilarities. 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 possibe.

2.1.3 Types of Clustering

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's produced in each iteration are subsets of a cluster of the previous iteration. This can occur by starting from one cluster, which contains all instances, and repeatidly partition the available clusters to even smaller ones.
Exclusive clustering signifies that an instance can only be a part of one and only cluster. Overlapping or non exclusive clustering signifies that an instance can be part of more than one clusters. Fuzzy clystering 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 cendroids 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 defintion from statistics as follows:
–TO DO–

Chapter 3

CASE STUDY: CLUSTER ANALYSIS ON TRAVEL AGENCY'S DATA

Bibliography

Bishop, C. M. (2006). *Pattern recognition and machine learning.*
Dunham, M. H. (2002). *Data mining: Introductory and advanced topics.*
Tan, P.-N., Kumar, V., & Steinbach, M. (2005). *Introduction to data mining.*