LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

FAKULTÄT FÜR MATHEMATIK, INFORMATIK UND STATISTIK
INSTITUT FÜR INFORMATIK

FORSCHUNGSGRUPPE
DATA MINING IN DER MEDIZIN

Bachelor Thesis
in Computer Science

Attention in Mixed-Type Clustering

BEARBEITERNAME

Aufgabensteller:	Prof. Dr. Christian Böhm
Betreuer:	BETREUERNAME
Abgabedatum:	tt.mm.yyyy

Declaration of Authorship

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references.
This paper was not previously presented to another examination board and has not been published.

Munich, tt.mm.yyyy

.....
BEARBEITERNAME

Abstract

This document serves as a model for the development of a thesis at the Department of Database Systems at the Institute for Computer Science at the LMU Munich. The abstract should not contain more than 300 words.

Contents

Chapter 1

Introduction

1.1 Clustering

Clustering is an unsupervised machine learning method that groups similar observations together. Due to its ability to find patterns in an unlabeled dataset, its an essential Task in Data Mining and Knowledge Discovery. A *cluster* is a group of similar observations that belong to a *centroid* (center point of a cluster). Distance-based clustering algorithms use distance measures such as Euclidean distance to calculate the similarity of datapoints. Hierarchical methods partition the observations and merge (agglomerative) or split them into bigger or smaller clusters. Many other methods exist, but this work focuses on methods for clustering *mixed-type* data. [1]

1.1.1 k-means

The most well known distance-based clustering method is k-means [3]. The goal is defined as follows: Suppose we have a finite set of n observations $S = \{p_1, p_2, ..., p_n\} \in \mathbb{R}^m$ for a dataset with m features, the target of k-means is to find $k(\leq n)$ optimal centroids $B = \{b_1, b_2, ..., b_k\} \subseteq \mathbb{R}^m$ that minimize the sum of the squared Euclidean distance of each point in S to its nearest centroid. Formally

$$\sum_{i=1}^{n} d(p_i, B)$$

has to be minimized, where d is the Euclidean distance from a point $p_i \in S$ to the nearest centroid in B; $d(p_i, B) = min_{1 \leq j \leq k} d(p_i, b_j)$ [4]. The Euclidean distance between two points p and $q$in an n-dimensional Euclidean space is defined as

$$d(p, q) = ||p - q|| = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2}$$

Finding the optimal centroids is a NP-hard problem, even for $d = 2$, as shown by Mahajan et al. [4]. The most common algorithm used for the k-means problem is a iterative refinement technique proposed by Lloyd [2] as follows:

1. Randomly set k initial cluster centroids $b_1^{(1)}, ..., b_k^{(1)}$.
2. Assign each obseration p_i to the nearest centroid using squared Euclidean distance. This splits our observations into S into k sets $\{S_1^{(t)}, ..., S_k^{(t)}\}$.
3. Recalculate the optimal position of each centroid using the mean distance to each observation assigned to the centroid.

$$b_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{p_j \in S_i^{(t)}} p_j$$

4. Repeat steps 2. and 3. until the centroid assignments no longer change.

Bibliography

[1] Amir Ahmad and Shehroz S. Khan. Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access*, 7:31883–31902, 2019.

[2] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[3] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[4] Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is np-hard. *Theoretical Computer Science*, 442:13–21, 2012. Special Issue on the Workshop on Algorithms and Computation (WALCOM 2009).