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

Jaanis Fehling

Aufgabensteller:	Prof. Dr. Christian Böhm
Betreuer:	Walid Durani
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

. . . .
Jaanis Fehling

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 Introduction

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 belongs 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]

1.2 k-means

The most well known distance-based clustering method is k-means [4]. 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; $d(p_i, B) = min_{1 \leq j \leq k} d(p_i, b_j)$ [5]. 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. [5]. The most common algorithm used for the k-means problem is a iterative refinement technique proposed by Lloyd [3]. It is defined 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.

1.3 Mixed-type data

In many real world scenarios, besides continuous, numerical data, *categorial* data exists. While Euclidean distance or other distance measures work well with continuous data, categorial data is different. Suppose we have states $\{A, B, C\}$ of a given feature, we would encode them into numeric values to allow for computation of a distance measure:

$$\{A, B, C\} \equiv \{1, 2, 3\}$$

While A and C can share the same semantic similarity as A and B, numerically category A and Category C are now $|1-3| = 2$ apart, while Category A and B are only $|1-2| = 1$ apart. During clustering, this could lead to observations being assigned to centroids based on a wrong distance assumption.

A possible solution is to use *one-hot encoding*, also known as *dummy coding* in classical statistics. One-hot encoding turns a discrete feature containing k mutually exclusive states into a vector x of length k, in which only one of the elements x_k equals 1 and all remaining elements equal 0 [2]. For an observation B of a feature having $k = 3$ separate states $\{A, B, C\}$, the one-hot vector x would be represented by $x = (0, 1, 0)^\intercal$.

	Naive k-means	k-means one-hot	k-prototypes	Gower distance
Soybean Disease	0.679950	0.674068	**0.694658**	0.669526
Heart Disease	**0.198871**	0.193008	0.153387	0.140792
Breast Cancer	**0.748214**	0.734909	0.587960	0.537113
Bank Marketing	0.017640	**0.029666**	0.017566	0.000356
Census Income	0.072767	**0.146271**	0.022475	0.000507
Credit Approval	**0.225376**	0.161375	0.114664	0.003465

Figure 1.1: Comparsion of Normalized Mutual Information (NMI) of classical methods on clustering mixed-type data. Naive k-means and k-means with one-hot encoding where calculated 100 times and the average NMI was taken. After calculating the Gower distance matrix, Agglomerative Clustering with average linkage between sets of observation was used. Due to high runtime and memory consumption, the "Bank Marketing" and "Census Income" datasets where downsampled to the first 2000 instances.

1.3.1 k-modes

1.4 Methodology

In this work we use 6 popular mixed-type datasets from the UC Irvine Machine Learning Repository [6].

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] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

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

[4] 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.

[5] 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).

[6] Kolby Nottingham Markelle Kelly, Rachel Longjohn. The UCI Machine Learning Repository. `https://archive.ics.uci.edu`. Accessed: 2023-07-16.