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 [5]. 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 optimal centroids $B = \{b_1, b_2, ..., b_k\} \subseteq \mathbb{R}^m$ for a given $k(\leq n) \in \mathbb{N}$ 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 [6]:

$$d(p_i, B) = min_{1 \leq j \leq k} d(p_i, b_j)$$

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. [6]. The most common algorithm used for the k-means problem is a iterative refinement technique proposed by Lloyd [4]. 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 categories $\{A, B, C\}$ 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 categories 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 categories $\{A, B, C\}$, the one-hot vector x would be represented by $x = (0, 1, 0)^\intercal$.

1.3.1 k-modes

According to Huang [3], one-hot encoding has two drawbacks:

1. In real-world applications, categorial features with hundreds or thousands of categories are encountered. This would result in a large number of binary features in the one-hot encoded representation, which will increase cost and space of computation.

2. The centroid value of a certain one-hot encoded feature, given by a real value between 0 and 1, cannot indicate the characteristics of the according cluster, since the feature only describes the presence or absence of one category.

Therefore, Huang [3] proposed using the Kronecker-Delta as a dissimilarity measure between multiple categorial columns. Formally, if we have two observations X and Y of a dataset with m categorial features, d_1 will count the number of mismatches between the categorial features of both instances, defined as

$$d_1(X, Y) = \sum_{j=1}^{m} \delta(x_j, y_j)$$

where the Kronecker delta $\delta(x_j, y_j)$ is defined as

$$\delta(x_j, y_j) = \begin{cases} 0, & (x_j = y_j) \\ 1, & (x_j \neq y_j) \end{cases}$$

If we have a finite set of n observations $S = \{p_1, p_2, ..., p_n\} \in \mathbb{R}^m$ for a dataset with m categorial features, the goal of *k-modes* [3] is to find optimal *modes* $B = \{b_1, b_2, ..., b_k\} \subseteq \mathbb{R}^m$ for a given $k(\leq n) \in \mathbb{N}$ that minimize

$$\sum_{i=1}^{n} d_1(p_i, B)$$

where

$$d_1(p_i, B) = min_{1 \leq l \leq k} \sum_{j=1}^{m} \delta(p_{i,j}, b_{l,j})$$

Similar to k-means, we can use an iterative algorithm for efficient computation [3]:

1. Randonly choose k observations from the dataset as initial modes for the clusters.

2. Assign each observation to their nearest mode using the proposed dissimilarity measure one by one and update the mode of each cluster after each assignment.

3. Test if each observation still belongs to its assigned mode, i.e. if each observation is assigned to its nearest mode. If the observation would belong to a different mode, reassign the observation and update the modes of both clusters.

4. Repeat step 3. until the mode assignments no longer change.

1.3.2 k-prototypes

As proposed by Huang [3], it is straightforward to combine the k-means and k-modes algorithms into the *k-prototypes* algorithm, which can be used to cluster *mixed-type* data (consisting of numerical, continuous and categorial features). The dissimilarity between two observations X and Y with features $A_1^r, A_2^r, ..., A_s^r, A_{s+1}^c, ..., A_m^c$, where features $A_1^r, ..., A_s^r$ are continuous and features $A_{s+1}^c, ..., A_m^c$ are categorial, is defined as

$$d_2(X,Y) = \sum_{j=1}^{s}(x_j - y_j)^2 + \gamma \sum_{j=s+1}^{m} \delta(x_j, y_j)$$

The first part of the equation is the Euclidean distance as used in k-means, while the second part is taken from the k-modes algorithm. Huang [3] states: "The weight γ is used to avoid favouring either type of attribute".

Again, we need to find k optimal centroids $B = \{b_1, b_2, ..., b_k\}$ and therefore have to minimize

$$\sum_{i=1}^{n} d_2(p_i, B)$$

where

$$d_2(p_i, B) = min_{1 \leq l \leq k} \sum_{j=1}^{s}(p_{i,j} - b_{l,j})^2 + \gamma \sum_{j=s+1}^{m} \delta(p_{i,j}, b_{l,j})$$

We can minimize both distance measures at the same time since they are nonnegative. Therefore, we can use the same algorithm as defined in 1.3.1. [3]

1.4 Methodology

In this work we use 8 mixed-type datasets from the UC Irvine Machine Learning Repository [7]. The Abalone dataset [9] contains physical measurements from abalones. It has 4177 observations, one categorial feature and seven continuous features. The Auction Verification dataset [10] has 2043 observations that contain verification runs of multi-round auctions. It is composed of six categorial and one continuous feauture. The Bank Marketing dataset [8] is related to a direct marketing campaign of a portuguese banking institution. It has 49732 observations, eight categorial and five continuous features.

All observations containing missing values were removed. Categorial columns were standardized by removing the mean and scaling to unit variance, using the scikit-learn Python API [11]. The standardized score z of a sample x from a feature is calculated as

$$z = \frac{(x - \mu)}{\sigma}$$

where μ is the mean of the samples $x_1, ..., x_N$ from a feature of length N, defined as

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

and σ is the standard deviation of the samples of a feature, defined as

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$

The datasets were shuffled. For all random operations, a random state of Integer Value 1 was used to ensure reproducibility.

1.5 Comparison of classical Clustering methods

	Naive k-means	k-means one-hot	k-prototypes	Gower distance
Abalone	0.171795	**0.173982**	0.171639	0.161416
Auction Verification	**0.016172**	0.007087	0.007667	0.006170
Bank Marketing	0.019781	**0.026060**	0.019522	0.001334
Breast Cancer	**0.746818**	0.736310	0.592480	0.553707
Census Income	0.108029	**0.184979**	0.141737	0.004259
Credit Approval	**0.313076**	0.171038	0.116579	0.003465
Heart Disease	**0.204577**	0.164486	0.189264	0.140792
Soybean Disease	0.672229	**0.710164**	0.567635	0.669526

Figure 1.1: Comparsion of Normalized Mutual Information of various classical methods on clustering mixed-type datasets.

	Naive k-means	k-means one-hot	k-prototypes	Gower distance
Abalone	0.135265	0.131434	0.134307	**0.195356**
Auction Verification	0.664709	0.576114	0.580519	**0.800783**
Bank Marketing	0.779600	0.786600	0.787200	**0.884200**
Breast Cancer	**0.960469**	0.950220	0.915081	0.900439
Census Income	0.608200	0.697600	0.625600	**0.768400**
Credit Approval	**0.808576**	0.705972	0.666156	0.548239
Heart Disease	0.334448	0.321070	0.424749	**0.565217**
Soybean Disease	0.576512	**0.599644**	0.471530	0.501779

Figure 1.2: Comparsion of Cluster Accuracy of various classical methods on clustering mixed-type datasets.

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] Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2:283–304, 1998.

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

[5] 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.

[6] 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).

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

[8] S. Moro, P. Rita, and P. Cortez. Bank Marketing. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C5K306.

[9] Warwick Nash, Tracy Sellers, Simon Talbot, Andrew Cawthorn, and Wes Ford. Abalone. UCI Machine Learning Repository, 1995. DOI: https://doi.org/10.24432/C55C7W.

[10] Elaheh Ordoni, Jakob Bach, Ann-Katrin Fleck, and Jakob Bach. Auction Verification. UCI Machine Learning Repository, 2022. DOI: https://doi.org/10.24432/C52K6N.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.