Pattern Recognition 47 (2014) 2505–2516

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

The MinMax *k*-Means clustering algorithm

Grigorios Tzortzis*, Aristidis Likas

Department of Computer Science & Engineering, University of Ioannina, Ioannina 45110, Greece

ARTICLE INFO

Article history:
Received 18 February 2013
Received in revised form
20 January 2014
Accepted 28 January 2014
Available online 6 February 2014

Keywords:
Clustering
k-Means
k-Means initialization
Balanced clusters

ABSTRACT

Applying *k*-Means to minimize the sum of the intra-cluster variances is the most popular clustering approach. However, after a bad initialization, poor local optima can be easily obtained. To tackle the initialization problem of *k*-Means, we propose the MinMax *k*-Means algorithm, a method that assigns weights to the clusters relative to their variance and optimizes a weighted version of the *k*-Means objective. Weights are learned together with the cluster assignments, through an iterative procedure. The proposed weighting scheme limits the emergence of large variance clusters and allows high quality solutions to be systematically uncovered, irrespective of the initialization. Experiments verify the effectiveness of our approach and its robustness over bad initializations, as it compares favorably to both *k*-Means and other methods from the literature that consider the *k*-Means initialization problem.

1. Introduction

Clustering is a fundamental problem in data analysis that arises in a variety of fields, such as pattern recognition, machine learning, bioinformatics and image processing [1,2]. It aims at partitioning a set of instances into homogeneous groups, i.e. the intra-cluster similarities are high while the inter-cluster similarities are low, according to some clustering objective. However, exhaustively searching for the optimal assignment of instances to clusters is computationally infeasible and usually a good local optimum of the clustering objective is sought.

One of the most well-studied clustering algorithms is *k*-Means [3], which minimizes the sum of the intra-cluster variances. Its simplicity and efficiency have established it as a popular means for performing clustering across different disciplines. Even an extension to kernel space has been developed [4,5] to enable the identification of nonlinearly separable groups. Despite its wide acceptance, *k*-Means suffers from a serious limitation. Its solution heavily depends on the initial positions of the cluster centers, thus after a bad initialization it easily gets trapped in poor local minima [6,7]. To alleviate this shortcoming, *k*-Means with multiple random restarts is often employed in practice.

Several methods attempt to overcome the sensitivity to the initialization in a more principled way. A first group of methods applies special techniques aiming at systematically avoiding partitionings of poor quality during the restarts. In [8], the initial centers are selected through a stochastic procedure such that the entire data space is covered. Theoretical guarantees are provided about the capability of the method to approximate the optimal clustering. Two approaches that start from random centers and penalize clusters relative to the winning frequency of their representatives are presented in [9,10]. Discouraging clusters to which several points have already been assigned from attracting new points in the subsequent steps has a regularizing effect. Centers that were initially ill-placed and are currently underutilized can actively participate in the solution on the following steps, which obstructs outlier clusters from forming and in effect balances the sizes of the clusters. Some other, analogous, strategies can be found in [11,12].

A second group of methods attempts to eliminate the dependence on random initial conditions, hence restarts are not anymore necessary. Global *k*-Means [13] and its modifications [14,15] are incremental approaches that start from a single cluster and at each step a new cluster is deterministically added to the solution according to an appropriate criterion. A kernel-based version of global *k*-Means is also available [16,17]. In [18] and its extension [19], spectral clustering is applied to locate the global optimum of a relaxed version of the *k*-Means objective, by formulating the problem as a trace maximization. Although these algorithms are not susceptible to bad initializations, they are computationally more expensive.

In this paper we propose MinMax *k*-Means, a novel approach that tackles the *k*-Means initialization problem by altering its objective. Our method starts from a randomly picked set of centers and tries to minimize the maximum intra-cluster variance instead of the sum of the intra-cluster variances. Specifically, *a weight is associated with each cluster*, such that clusters with larger variance[1] are allocated higher weights, and a weighted version of the sum of the intra-cluster variances criterion is derived. Different notions of weights have been

* Corresponding author. Tel.: +30 26510 08838; fax: +30 26510 08882.
E-mail addresses: gtzortzi@cs.uoi.gr (G. Tzortzis), arly@cs.uoi.gr (A. Likas).

[1] To avoid cluttering the text, we shall also refer to the intra-cluster variances, simply, as the variances of the clusters.

http://dx.doi.org/10.1016/j.patcog.2014.01.015
0031-3203 © 2014 Elsevier Ltd. All rights reserved.

exploited in the literature across several *k*-Means variants. In fuzzy *c*-means and Gaussian mixture models [20] weights are used to compute the degree of cluster membership of the instances, while in other variants weights are assigned to features, or groups of features, such that the tasks of clustering and feature selection are simultaneously performed [21,22]. Also, in [23], a weighting factor is added to each instance in order to detect outliers.

The per cluster weights predispose our algorithm towards primarily minimizing those clusters that currently exhibit a large variance, in essence confining the occurrence of large variance clusters in the outcome, and are *learned automatically*, together with the cluster assignments. The proposed method alternates between a minimization step, resembling the *k*-Means procedure, and an additional maximization step, in which the weights are calculated using closed-form expressions. By applying this weighting mechanism, results become less affected by the initialization and *solutions of high quality can be more consistently discovered*, even after starting from a bad initial set of centers. In addition, the obtained clusters are balanced with respect to their variance.

The presented algorithm also *incorporates a parameter p*, whose value must be specified prior to execution, that adjusts the degree of its bias towards penalizing large variance clusters. When $p=0$, *k*-Means, which has a zero bias, can be deduced as a special case of our method. A practical framework extending MinMax *k*-Means to automatically adapt this parameter to the dataset has been also developed in this work, so that the hidden group structures in the data can be successfully uncovered.

Experiments are conducted on several diverse datasets, including images, handwritten digits, proteins and patient records. MinMax *k*-Means is compared to *k*-Means, as well as to *k*-Means++ [8] and pifs *k*-Means [10] that evade degenerate optima, the first by methodically picking the initial centers and the second by balancing the cluster sizes. Our empirical evaluation reveals the effectiveness of the proposed clustering scheme in restricting the emergence of large variance clusters and producing superior solutions compared to the other three approaches, while restarted from random initializations. Furthermore, we observe that our algorithm constitutes a very promising technique for initializing *k*-Means.

The rest of this paper is organized as follows. We next briefly describe *k*-Means, while in Section 3 the proposed MinMax *k*-Means algorithm is presented and its properties are analyzed. Section 4 introduces our practical framework for setting the *p* parameter. The experiments follow in Section 5, before the concluding remarks of Section 6.

2. *k*-Means

To partition a dataset $\mathcal{X}=\{\mathbf{x}_i\}_{i=1}^N$, $\mathbf{x}_i \in \Re^d$ into M disjoint clusters, $\{\mathcal{C}_k\}_{k=1}^M$, *k*-Means [3] minimizes the sum of the intra-cluster variances (1), where $\mathcal{V}_k=\sum_{i=1}^N \delta_{ik}\|\mathbf{x}_i-\mathbf{m}_k\|^2$ and $\mathbf{m}_k=\sum_{i=1}^N \delta_{ik}\mathbf{x}_i/\sum_{i=1}^N \delta_{ik}$ are the variance[2] and the center of the *k*-th cluster, respectively, and δ_{ik} is a cluster indicator variable with $\delta_{ik}=1$ if $\mathbf{x}_i \in \mathcal{C}_k$ and 0 otherwise.

$$\mathcal{E}_{sum}=\sum_{k=1}^M \mathcal{V}_k=\sum_{k=1}^M\sum_{i=1}^N \delta_{ik}\|\mathbf{x}_i-\mathbf{m}_k\|^2 \tag{1}$$

Clustering proceeds by alternating between assigning instances to their closest center and recomputing the centers, until a local minimum is (monotonically) reached.

Despite its simplicity and speed, *k*-Means has some drawbacks, with the most prominent being the dependence of the solution on the choice of initial centers [6,7]. Bad initializations can lead to poor local minima, thus multiple random restarts are usually executed to circumvent the initialization problem. Often, the solutions returned by the restarts significantly vary in terms of the achieved objective value, ranging from good to very bad ones, particularly for problems with a large search space (e.g. many clusters and dimensions). Therefore, numerous runs of the algorithm are required to increase the possibility of locating a good local minimum.

3. MinMax *k*-Means

As discussed in Section 2, the sensitivity of *k*-Means to initialization and the diverse solutions uncovered during the restarts make it difficult to find a good partitioning of the data. Motivated by this, we propose the optimization of a different objective and a new methodology that allows *k*-Means to *produce high quality partitionings more systematically*, while restarted from random initial centers.

3.1. *The maximum variance objective*

Consider a dataset $\mathcal{X}=\{\mathbf{x}_i\}_{i=1}^N$, $\mathbf{x}_i \in \Re^d$ to be split into M disjoint clusters, $\{\mathcal{C}_k\}_{k=1}^M$. Instead of minimizing the *sum* of the intra-cluster variances (1), we propose to minimize the *maximum* intra-cluster variance

$$\mathcal{E}_{max}=\max_{1\le k\le M}\mathcal{V}_k=\max_{1\le k\le M}\left\{\sum_{i=1}^N \delta_{ik}\|\mathbf{x}_i-\mathbf{m}_k\|^2\right\}, \tag{2}$$

where $\mathcal{V}_k$, $\mathbf{m}_k$ and δ_{ik} are defined as in (1).

The rationale for this approach is the following: the summation over all clusters in the *k*-Means objective (1) allows for similar $\mathcal{E}_{sum}$ values to be achieved either by having a few clusters with large variance that are counterbalanced by others with small variance, or by having a moderate variance for all clusters. This means that the relative differences among the cluster variances are not taken into account. Note that the variance of a cluster is a measure of its quality. The above remark does not hold when minimizing $\mathcal{E}_{max}$ though, as the first case above would lead to a higher objective value. Hence, when minimizing $\mathcal{E}_{max}$, large variance clusters are avoided and *the solution space is now restricted towards clusters that exhibit more similar variances*.

The previous observation has two important implications. Since *k*-Means minimizes $\mathcal{E}_{sum}$, it cannot distinguish between the two cases, thus a bad initialization yields a poor solution that is characterized by substantially different variances among the returned clusters; a result of natural groups getting merged (large variance clusters) and others getting split (small variance clusters), or of outlier clusters being formed.[3] As explained, the maximum intra-cluster variance objective $\mathcal{E}_{max}$ is less likely to converge to such solutions, hence it is *easier to overcome a bad initialization*. Thus, we expect a *k*-Means type algorithm coupled with this objective to be able to uncover better group structures more consistently during the restarts. An example is illustrated in Fig. 1.

Additionally, a balancing effect on the clusters occurs. Balanced outcomes have been pursued in different ways in the literature. For example, in [10] *k*-Means and spherical *k*-Means are modified to penalize clusters in proportion to the number of instances assigned to them, while in [24,25] a graph cut criterion is optimized which favors the creation of subgraphs where the sums of the edge weights within the subgraphs (subgraph associations) are similar. In our case, balancing is done with regard to the

[2] In this work, we define cluster variance as the sum, and not the average, of the squared distances from the instances belonging to the cluster to its center.

[3] Let us clarify that a solution with quite different variances on the clusters is not necessarily a bad one. There are datasets where the natural groups exhibit such structure. We simply claim that such behavior also arises after a bad initialization, where some groups are merged and others are split.