Pattern Recognition Letters 45 (2014) 71–77

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Improving cluster analysis by co-initializations ☆

He Zhang *, Zhirong Yang, Erkki Oja

Department of Information and Computer Science, Aalto University, Espoo, Finland

ARTICLE INFO

Article history:
Received 11 April 2013
Available online 16 March 2014

Keywords:
Clustering
Initializations
Cluster ensembles

ABSTRACT

Many modern clustering methods employ a non-convex objective function and use iterative optimization algorithms to find local minima. Thus initialization of the algorithms is very important. Conventionally the starting guess of the iterations is randomly chosen; however, such a simple initialization often leads to poor clusterings. Here we propose a new method to improve cluster analysis by combining a set of clustering methods. Different from other aggregation approaches, which seek for consensus partitions, the participating methods in our method are used consequently, providing initializations for each other. We present a hierarchy, from simple to comprehensive, for different levels of such co-initializations. Extensive experimental results on real-world datasets show that a higher level of initialization often leads to better clusterings. Especially, the proposed strategy is more effective for complex clustering objectives such as our recent cluster analysis method by low-rank doubly stochastic matrix decomposition (called DCD). Empirical comparison with three ensemble clustering methods that seek consensus clusters confirms the superiority of improved DCD using co-initialization.

1. Introduction

Cluster analysis plays an essential role in machine learning and data mining. The aim of clustering is to group a set of objects in such a way that the objects in the same cluster are more similar to each other than to the objects in other clusters, according to a particular objective. Many clustering methods are based on objective functions which are non-convex. Their optimization generally involves iterative algorithms which start from an initial guess. Proper initialization is critical for getting good clusterings.

For simplicity, random initialization has been widely used, where a starting point is randomly drawn from a uniform or other distribution. However, such a simple initialization often yields poor results and the iterative clustering algorithm has to be run many times with different starting points in order to get better solutions. More clever initialization strategies are thus required to improve efficiency.

Many ad hoc initialization techniques have been proposed for specific clustering methods, for example, specific choices of the initial cluster centers of the classical *k*-means method (see e.g. [1–4]), or singular value decomposition for clustering based on nonnegative matrix factorization [5,6]. However, there seems to be no initialization principle that would be commonly applicable for a wide range of iterative clustering methods. Especially, there is little research on whether one clustering method can benefit from initializations by the results of another clustering method.

In this paper, we show experimentally that the clusterings can usually be improved if a set of diverse clustering methods provide initializations for each other. We name this approach *co-initialization*. We present a hierarchy of initializations towards this direction, where a higher level represents a more extensive strategy. At the top are two levels of co-initialization strategies. We point out that despite their extra computational cost, these strategies can often bring significantly enhanced clustering performance. The enhancement is especially significant for more complex clustering objectives, for example, Probabilistic Latent Semantic Indexing [7], and our recent clustering method by low-rank doubly stochastic matrix decomposition (called DCD) [8].

Our claims are supported by extensive experiments on nineteen real-world clustering tasks. We have used a variety of datasets from different domains such as text, vision, and biology. The proposed initialization hierarchy has been tested using eight state-of-the-art clustering methods. Two widely used criteria, cluster purity and Normalized Mutual Information, are used to measure the clustering performance. The experimental results verify that a higher level initialization in the proposed hierarchy often achieve better clustering performance.

Ensemble clustering is another way to combine a set of clustering methods. It aggregates the different clusterings into a single

☆ This paper has been recommended for acceptance by Andrea Torsello.

* Corresponding author. Tel.: +358 505188888.
E-mail addresses: he.zhang@aalto.fi (H. Zhang), zhirong.yang@aalto.fi (Z. Yang), erkki.oja@aalto.fi (E. Oja).

http://dx.doi.org/10.1016/j.patrec.2014.03.001
0167-8655/

one. We also compared co-initialization with three prominent ensemble clustering methods. The comparison results show that the improved DCD using co-initializations outperforms these ensemble approaches that seek a consensus clustering.

In the following, Section 2 reviews briefly the recently introduced Data-Cluster-Data (DCD) method. It is a representative clustering method among those that strongly benefit from co-initializations, and will be shown to be overall the best method in the experiments. Then Section 3 reviews related work on ensemble clustering, which is another way of combining a set of base clustering methods. In Section 4, we present our novel co-initialization method and describe the initialization hierarchy. Experimental settings and results are reported in Section 5. Section 6 concludes the paper and discusses potential future work.

2. Clustering by DCD

Some clustering methods such as Normalized Cut [9] are not sensitive to initializations but tend to return less accurate clustering (see e.g. [10], page 8, [8,11], and Section 5.3). On the other hand, some methods can find more accurate results but require careful initialization. The latter kind of methods can benefit more from our co-initialization strategy, to be introduced in Section 4. Recently we proposed a typical clustering method of the latter kind, which is based on Data-Cluster-Data random walk and thus called DCD [8]. In this section we recapitulate the essence of DCD. It belongs to the class of probabilistic clustering methods. Given n data samples and r clusters, denote by $P(k|i)$ the probability of assigning the ith sample to the kth cluster, where $i = 1,\ldots,n$ and $k = 1,\ldots,r$.

Suppose the similarities between data items are precomputed and given in an $n \times n$ nonnegative symmetric sparse matrix A. DCD seeks an approximation to A by another matrix $\widehat{A}$ whose elements correspond to the probabilities of two-step random walks between data points through clusters. Let i,j, and v be indices for data points, and k and l for clusters. Then the random walk probabilities are given as

$$\widehat{A}_{ij} = P(i|j) = \sum_k P(i|k)P(k|j) = \sum_k \frac{P(k|i)P(k|j)}{\sum_v P(k|v)}, \tag{1}$$

by using the Bayes formula and the uniform prior $P(i) = 1/n$.

The approximation is given by the Kullback–Leibler (KL-) divergence. This is formulated as the following optimization problem [8]:

$$\underset{P\geqslant 0}{\text{minimize}} \quad D_{\mathrm{KL}}(A||\widehat{A}) = \sum_{ij}\left(A_{ij}\log\frac{A_{ij}}{\widehat{A}_{ij}} - A_{ij} + \widehat{A}_{ij}\right), \tag{2}$$

where $\widehat{A}_{ij} = \sum_k \frac{P_{ik}P_{jk}}{\sum_v P_{vk}}$ with $P_{ik} = P(k|i)$, subject to $\sum_k P_{ik} = 1,\ i = 1,\ldots,n$.

Denote $\nabla = \nabla^+ - \nabla^-$ as the gradient of $D_{\mathrm{KL}}(A||\widehat{A})$ with respect to P, where ∇^+ and ∇^- are the positive and (unsigned) negative parts of ∇, respectively. The optimization is solved by a Majorization–Minimization algorithm [12–15] that iteratively applies a multiplicative update rule:

$$P_{ik} \leftarrow P_{ik}\frac{\nabla^-_{ik}a_i + 1}{\nabla^+_{ik}a_i + b_i}, \tag{3}$$

where $a_i = \sum_l \frac{P_{il}}{\nabla^+_{il}}$ and $b_i = \sum_l P_{il}\frac{\nabla^-_{il}}{\nabla^+_{il}}$.

The preprocessing of DCD employs the common approximation of making A sparse by zeroing the non-local similarities. This makes sense for two reasons: first, geodesics of curved manifolds in high-dimensional spaces can only be approximated by Euclidean distances in small neighborhoods; second, most popular distances computed of weak or noisy indicators are not reliable over long distances, and the similarity matrix is often approximated by the K-Nearest Neighbor graph with good results, especially when n is large. With a sparse A, the computational cost of DCD is $O(|E| \times r)$ for $|E|$ nonzero entries in A and r clusters. In the experiments we used symmetrized and binarized K-Nearest-Neighbor graph as A ($K \ll n$). Thus the computational cost is $O(nKr)$.

Given a good initial decomposing matrix P, DCD can achieve better cluster purity compared with several other state-of-the-art clustering approaches, especially for large-scale datasets where the data points situate in a curved manifold. Its success comes from three elements in its objective: (1) the approximation error measure by Kullback–Leibler divergence takes into account sparse similarities; (2) the decomposing matrix P as the only variable to be learned contains just enough parameters for clustering; and (3) the decomposition form ensures relatively balanced clusters and equal contribution of each data sample.

What remains is how to get a good starting point. The DCD optimization problem is harder to solve than conventional NMF-type methods based on Euclidean distance in three aspects: (1) the geometry of the KL-divergence cost function is more complex; (2) DCD employs a structural decomposition where P appears more than once in the approximation, and appears in both numerator and denominator; (3) each row of P is constrained to be in the $(r-1)$-simplex. Therefore, finding a satisfactory DCD solution requires more careful initialization. Otherwise the optimization algorithm can easily fall into a poor local minimum.

Yang and Oja [8] proposed to obtain the starting points by pre-training DCD with regularization term $(1-\alpha)\sum_{ik}\log P_{ik}$. This corresponds to imposing Dirichlet priors over the rows of P. By varying α, the pre-training can provide different starting points for multiple runs of DCD. The final result is given by the one with the smallest DCD objective of Eq. 2. This initialization strategy can bring improvement for certain datasets, whereas the enhancement remains mediocre as it is restricted to the same family of clustering methods. In the remaining, we investigate the possibility to obtain good starting points with the aid of other clustering methods.

3. Ensemble clustering

In supervised machine learning, it is known that combining a set of classifiers can produce better classification results (see e.g. [16]). There have been also research efforts with the same spirit in unsupervised learning, where several basic clusterings are combined into a single categorical output. The base results can come from results of several clustering methods, or the repeated runs of a single method with different initializations. In general, after obtaining the bases, a combining function, called *consensus function*, is needed for aggregating the clusterings into a single one. We call such aggregating methods *ensemble cluster analysis*.

Several ensemble clustering methods have been proposed. An early method [17] first transforms the base clusterings into a hypergraph and then uses a graph-partitioning algorithm to obtain the final clusters. Gionis et al. [18] defined the distance between two clusterings as the number of pairs of objects on which the two clusterings disagree, based on which they formulated the ensemble problem as the minimization of the total number of disagreements with all the given clusterings. Fred and Jain [19] explored the idea of evidence accumulation and proposed to summarize various clusterings in a co-association matrix. The incentive of their approach is to weight associations between sample pairs by the number of times they co-occur in a cluster from the set of given clusterings. After obtaining the co-association matrix, they applied the agglomerative clustering algorithm to yield the final partition. Iam-On et al. [20] introduced new methods for generating two