Graph-Optimized Low-Rank Doubly Stochastic Decomposition Clustering Postprint

Authors: Zhang Tao, Enliang Hu, Yu Jingli

Date: 2018-05-20T00:00:00+00:00

Abstract

Low-rank doubly stochastic matrix decomposition for cluster analysis (DCD) is a graph clustering method recently proposed by Yang et al. [16], which obtains a non-negative low-rank doubly stochastic matrix decomposition A=UUT(U(0) from the graph association matrix S by minimizing the KL (Kullback-Leibler) divergence criterion KL(A, S), and uses U as the class label matrix for clustering. In the DCD method, because matrix S is fixed and immutable, the quality of the initial value selection for S has a significant impact on the clustering results, leading to its lack of stability. To address this issue, a graph optimization-based DCD method is proposed that integrates the optimization of the graph association matrix S and DCD into a unified framework, thereby improving and extending the original DCD method. Experimental results demonstrate that, compared with the DCD method, the graph optimization-based DCD method achieves better clustering accuracy and stability.

Full Text

Graph-Optimized Low-Rank Doubly Stochastic Decomposition for Clustering

Zhang Tao, Hu Enliang†, Yu Jingli
(Department of Mathematics, Yunnan Normal University, Kunming, Yunnan 650500, China)

Abstract

Low-rank doubly stochastic matrix decomposition for cluster analysis (DCD) is a graph clustering method recently proposed by Yang et al. [**?**]. DCD obtains a nonnegative low-rank doubly stochastic decomposition $A = UU^T$ $(U \geq 0)$ from a graph affinity matrix S by minimizing the Kullback-Leibler (KL) divergence criterion $\mathrm{KL}(A, S)$, and uses U as the class label matrix for clustering. In DCD,

since matrix S is fixed and unchangeable, the initial value of S has a significant impact on clustering results, leading to instability. To address this issue, we propose a graph-optimized DCD method that integrates the optimization of graph affinity matrix S and DCD within a unified framework, thereby improving and extending the original DCD method. Experimental results demonstrate that the graph-optimized DCD achieves better clustering accuracy and stability compared to DCD.

Keywords: low-rank doubly stochastic matrix decomposition; graph optimization; stability; clustering

---

1 Introduction

Clustering is the process of grouping inherently unlabeled objects into different clusters based on the principle that "similar objects cluster together," with the goal of making objects within the same cluster similar to each other while sufficiently dissimilar to objects in other clusters. Clustering analysis constitutes an important research area in machine learning, data mining, and pattern recognition. Based on methodological types, clustering algorithms can be broadly categorized into: partition-based methods such as K-means [**?**] and K-medoids [**?**]; hierarchical methods such as CURE [**?**]; grid-based methods such as STING [**?**]; density-based methods such as DBSCAN [**?**]; neural network-based methods such as SOM [**?**]; and graph-based methods such as Normalized Cut [**?**]. While different clustering methods have their respective advantages, they also suffer from certain limitations to varying degrees, making the exploration of new clustering methods significant. The novel clustering method proposed in this paper belongs to the family of graph-based clustering approaches.

1.1 Graph Clustering

Graph clustering algorithms [**?**, **?**] are built upon graph theory. Their essence lies in first representing relationships between objects using a graph, then transforming the clustering problem into a graph partitioning problem—this constitutes a point-to-point clustering algorithm. In graph clustering, the graph structure among objects is expressed by an affinity matrix, and the quality of graph construction ultimately determines clustering performance. Graph construction typically involves two steps: edge selection and edge weight configuration. Widely used edge construction methods include K-nearest neighbor graphs [**?**], ϵ-ball neighbor graphs [**?**], and fully connected graphs. After graph edges are established, various edge weight configuration [**?**] methods exist, among which the most commonly used are 0-1 binary weights and weight settings using heat kernel functions [**?**].

1.2 Low-Rank Doubly Stochastic Matrix Decomposition Clustering

Over the past decade, low-rank matrix decomposition techniques have gradually found numerous applications in machine learning and data mining. In particular, nonnegative low-rank matrix decomposition has been successfully applied to clustering. In 1999, Hofmann [**?**] proposed using probabilistic latent semantic indexing for data segmentation, where KL divergence replaced traditional Euclidean distance in matrix decomposition. In 2001, Lee et al. [**?**] proposed nonnegative matrix factorization that decomposes a matrix into the product of two nonnegative low-rank matrices. In 2010, Ding et al. [**?**] demonstrated that nonnegative matrix factorization approximates traditional K-means methods. In 2013, Arora et al. [**?**] proposed that left stochastic matrix decomposition approximates similarity matrices generated by left stochastic matrices. Recently, Yang et al. [**?**] introduced a graph clustering method based on low-rank doubly stochastic matrix decomposition (DCD). The main idea of DCD is to minimize the KL divergence between a graph affinity matrix and a low-rank doubly stochastic matrix, where the doubly stochastic matrix is formed by the product of a clustering label matrix.

If we denote $\text{rank}(U) = r$, the set of r-rank doubly stochastic matrices can be expressed as follows. If we denote $\text{rank}(W) = r$ and $W \geq 0$, Yang et al. [**?**] proved that the above set is equivalent to $\mathcal{B}$, leading to the following theorem.

Theorem 1 [**?**]
This theorem demonstrates that $\mathcal{B}$ is also a set of doubly stochastic matrices. Compared to the representation of set $\mathcal{A}$, the representation of set $\mathcal{B}$ is more conducive to optimization. Therefore, subsequent descriptions of doubly stochastic matrix sets in this paper will be based on set $\mathcal{B}$.

1.3 Limitations of DCD and Proposed Improvements

In graph clustering methods, graph construction is unsupervised and thus involves certain randomness, leading to several shortcomings: (a) The graph or its corresponding affinity matrix S_0 is manually predefined and cannot be optimized during subsequent learning; (b) Graph construction only utilizes the spatial structure of raw data, which may not be most beneficial for subsequent clustering tasks; (c) Graph construction involves edge weight configuration methods, often leading to parameter selection difficulties (e.g., kernel parameter selection when using heat kernel weights).

To address these limitations, inspired by graph optimization dimensionality reduction research [**?**, **?**], we propose Graph-optimized low-rank Doubly stochastic matrix decomposition for Clustering (GoDCD) in Section 2. This method integrates the graph optimization process into the DCD objective function optimization, thereby achieving simultaneous learning of graph (affinity matrix) optimization and doubly stochastic matrix decomposition. The advantage of our algorithm is that in GoDCD, graph construction is not initially fixed but is gradually optimized during algorithm iterations, thus reducing dependence on

the initial affinity matrix and finding a more suitable graph affinity matrix for subsequent clustering tasks.

---

2 Graph-Optimized Doubly Stochastic Decomposition Clustering

2.1 Model Formulation

In the DCD model, graph construction is equivalent to constructing an initial graph affinity matrix S_0. If S_0 is poorly constructed, subsequent clustering performance will suffer. To partially overcome this problem, the GoDCD model proposed in this paper integrates graph optimization with the DCD clustering model into a unified learning framework. Its objective function is:

$$\min_{W,S} J(W,S) = \mathrm{KL}(S,B) + \alpha \sum_{i,j} (\log S_{ij} - \log S_{0ij}) + \lambda \mathrm{KL}(S,S_0) \quad (2)$$

where $B = WW^T$ with $W \geq 0$, $\sum_k W_{ik} = 1$, and $\sum_k W_{kj} = 1$.

Comparing the two models in equations (1) and (2), we can easily observe the differences between GoDCD and DCD:

(a) GoDCD has an additional term $\mathrm{KL}(S,S_0)$ compared to DCD, whose role is to optimize a better affinity matrix S within the neighborhood of S_0;

(b) The term $\mathrm{KL}(B,S_0)$ in DCD is replaced by $\mathrm{KL}(B,S)$ in GoDCD, with the purpose of performing low-rank doubly stochastic decomposition clustering based on a better affinity matrix S (rather than S_0);

(c) In DCD, only W is optimized, whereas in GoDCD both W and S are optimized simultaneously, which is equivalent to integrating graph optimization and low-rank doubly stochastic decomposition into the same objective function. The goal is to achieve joint optimality of graph construction (corresponding to S) and clustering (corresponding to W).

2.2 Model Solution

Since the objective function $J(W,S)$ is non-convex, solving problem (2) constitutes a non-convex optimization problem. For this problem, we employ an alternating minimization method for iterative solution: first fix S and solve the subproblem with respect to W; then fix W and solve the subproblem with respect to S:

$$W^{(t)} = \arg\min_{W} J(W, S^{(t-1)}) \quad (3)$$

$$S^{(t)} = \arg\min_{S} J(W^{(t)}, S) \quad (4)$$

For solving subproblem (3), we can directly use the DCD solution algorithm from [**?**]. For subproblem (4), its solution has a closed form. The specific derivation is as follows:

Taking the partial derivative of J with respect to S_{ij}, we obtain:

$$\frac{\partial J}{\partial S_{ij}} = \log S_{ij} - \log B_{ij} + \lambda(\log S_{ij} - \log S_{0ij}) = 0$$

This yields the iterative sequence:

$$S_{ij}^{(t+1)} = S_{ij}^{(t)} \times \left(\frac{B_{ij}^{(t)}}{S_{ij}^{(t)}} \right)^{\frac{1}{1+\lambda}} \quad (5)$$

Algorithm 1: GoDCD Solution Algorithm

Input: Initial affinity matrix S_0, number of clusters r, parameters α, λ
Output: Clustering label matrix W

1. Initialize: $S^{(0)} = S_0$, $t = 1$
2. **Repeat**
   - **Step 1 (Update W):** Solve subproblem (3) using the DCD algorithm to obtain $W^{(t)}$
   - **Step 2 (Update S):** Compute $S^{(t+1)}$ using equation (5)
   - **Step 3:** $t = t + 1$
3. **Until** $|J(W^{(t)}, S^{(t)}) - J(W^{(t-1)}, S^{(t-1)})| \leq \epsilon$ or $t >$ itermax

Theorem 2 (Convergence)
If $\{W^{(t)}, S^{(t)}\}$ is the sequence generated by Algorithm 1, then the sequence $\{J(W^{(t)}, S^{(t)})\}$ is monotonically decreasing. Moreover, since $J(W, S) \geq 0$, the iterative sequence is bounded below. According to the monotone convergence theorem, a bounded monotone decreasing sequence must have a limit, therefore $\{J(W^{(t)}, S^{(t)})\}$ has a limit, which proves that Algorithm 1 converges.

---

3 Experimental Results and Analysis

3.1 Data Description and Experimental Settings

In our experiments, we use fully-connected heat kernel weight graphs as the basis, with the corresponding initial graph affinity matrix S_0 where $S_{0ij} = \exp(-\|x_i - x_j\|^2/2\sigma^2)$ represents the affinity between x_i and x_j. We select nine datasets for experiments: iris, leaf4, sonar, chessboard, wine,

glass, heart, Balance_scale, and breast_cancer, all from the UCI repository (http://archive.ics.uci.edu/ml/datasets.html). Their information is shown in .

We compare three methods: Ncut [**?**], DCD [**?**], and our proposed GoDCD method. The evaluation metric is clustering purity [**?**], defined as:

$$\mathrm{CP} = \frac{1}{n}\sum_{k=1}^{r}\max_{1\le l\le r} n_{kl}$$

where n is the total number of samples in the dataset, and n_{kl} is the number of samples that belong to class l in the original data but are assigned to cluster k after algorithm clustering (i.e., $n_{kl} = |\{x_i \mid \mathrm{cluster}(x_i) = k, \mathrm{label}(x_i) = l\}|$). The clustering purity ranges between $[0, 1]$, with larger values indicating higher clustering accuracy.

3.2 Comparison of Clustering Accuracy

shows the clustering purity comparison results. Clustering purity measures the agreement between clustering labels and ground-truth labels. To validate the effectiveness of GoDCD, we present clustering purity comparisons in Table 2, from which we observe:

(a) DCD achieves significantly higher clustering purity than Ncut on iris, sonar, wine, chessboard, and heart datasets. This is because, compared to Ncut, DCD not only considers and utilizes the graph structure of data but also employs low-rank doubly stochastic decomposition to enhance clustering performance.

(b) Except on chessboard and heart datasets, GoDCD achieves notably higher clustering purity than DCD. Particularly on the iris dataset, GoDCD outperforms DCD by approximately 20%. The reason is that DCD only considers clustering optimality on the initial graph construction, whereas GoDCD simultaneously considers joint optimality of both graph construction and clustering.

(c) On the leaf4 dataset, both GoDCD and the original DCD method perform worse than Ncut, possibly indicating that doubly stochastic decomposition methods are not suitable for this dataset.

3.3 Influence of Model Parameters

3.3.1 Effect of Parameter α Parameter α is an additional model parameter in GoDCD relative to DCD. If α is too large, clustering becomes heavily dependent on the initial affinity matrix; if α is too small, the updated affinity matrix may deviate too far from the original data structure. However, selecting optimal α belongs to model selection problems, for which no reliable theory currently exists. [Figure 3: see original paper] and [Figure 4: see original paper] show clustering purity on heart and iris datasets using different α values. We

observe that when α varies within a certain range, clustering purity changes relatively smoothly, partially demonstrating that GoDCD is relatively stable with respect to parameter α. One reason is that even if the initial affinity matrix is not well-chosen, its optimization in GoDCD reduces dependence on the initial graph construction.

3.3.2 Effect of Parameter λ If clustering purity fluctuates only slightly across different values of parameter λ, the algorithm is considered relatively stable with respect to λ. [Figure 1: see original paper] and [Figure 2: see original paper] present clustering purity on wine and iris datasets using different λ values for both DCD and GoDCD. [Figure 1: see original paper] shows that compared to DCD, GoDCD exhibits smaller accuracy fluctuations across different λ values, indicating that GoDCD is more stable than DCD.

---

4 Conclusion

To improve clustering performance, this paper proposes Graph-optimized low-rank Doubly stochastic matrix decomposition for Clustering (GoDCD), which generalizes the original DCD clustering method. In GoDCD, graph optimization and low-rank doubly stochastic decomposition clustering are integrated into a unified objective function. The role of this integration is to achieve joint optimality of graph construction and clustering, thereby reducing the dependence of subsequent clustering on the quality of initial graph construction. Experimental results on several UCI datasets demonstrate that in most cases, GoDCD achieves higher clustering accuracy and better stability than DCD.

Currently, the GoDCD method is only applied to unsupervised clustering problems. However, effective semi-supervised auxiliary information would help achieve more accurate clustering. Therefore, extending GoDCD to semi-supervised clustering scenarios represents a worthwhile direction for future research.

References

[1] Hornik K, Feinerer I, Kober M, et al. Spherical K-means clustering [J]. Journal of Statistical Software, 2012, 50(10): 1-22.

[2] Park H S, Jun C H. A simple and fast algorithm for K-medoids clustering [J]. Expert Systems with Applications, 2009, 36(2): 3336-3341.

[3] Wang W Y, Wang C X, Wang J. Research on Hybrid parallel programming technique based on cmp multi-cure cluster [J]. Computer Science, 2014, 41(2): 19-22.

[4] Festing M, Royer S, Steffen C. Do clusters help firms to realise competitive advantage? A resource-based analysis of the mechanical watch cluster in Glashütte//Germany [J]. Zeitschrift Für Management, 2010, 5(2): 165-185.

[5] Pan D, Zhao L. Uncertain data cluster based on DBSCAN [C]//Proc of International Conference on Multimedia Technology. 2011: 3781-3784.

[6] Samsonova E V, Kok J N, Ijzerman A P. TreeSOM: cluster analysis in the self-organizing map [J]. Neural Networks the Official Journal of the International Neural Network Society, 2006, 19(6-7): 935.

[7] Lagrange M, Martins L G, Murdoch J, et al. Normalized cuts for predominant melodic source separation [J]. IEEE Trans on Audio Speech & Language Processing, 2008, 16(2): 278-290.

[8] 李建元, 周脚根, 关佶红, 等. 谱图聚类算法研究进展 [J]. 智能系统学报, 2011, 06(5): 405-414.

[9] Luxburg U V. A tutorial on spectral clustering [J]. Statistics and Computing, 2007, 17(4): 395-416.

[10] Arya S, Malamatos T, Mount D M. Space-time tradeoffs for approximate nearest neighbor searching [J]. Journal of the Acm, 2009, 57(1): 1-54.

[11] He X, Niyogi P. Locality preserving projections [J]. Advances in Neural Information Processing Systems, 2003, 16(1): 186-197.

[12] Belkin, Mikhail, Niyogi, et al. Laplacian Eigenmaps for dimensionality reduction and data representation [J]. Neural Computation, 2014, 15(6): 1373-1396.

[13] Maier M, Luxburg U V, Heiny M. Influence of graph construction on graph-based clustering measures [J]. Nips, 2009(2009): 1025-1032.

[14] Hofmann T. Probabilistic latent semantic indexing [C]//Proc of International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1999: 50-57.

[15] Lee D D, Seung H S. Algorithms for non-negative matrix factorization [C]//Proc of International Conference on Neural Information Processing Systems. MIT Press, 2000: 535-541.

[16] Yang Z, Corander J, Oja E. Low-rank doubly stochastic matrix decomposition for cluster analysis [J]. Journal of Machine Learning Research, 2016, 17(1): 6454-6478.

[17] Zhang L, Chen S, Qiao L. Graph optimization for dimensionality reduction with sparsity constraints [J]. Pattern Recognition, 2012, 45(3): 1205-1210.

[18] Zhang L, Qiao L, Chen S. Graph-optimized locality preserving projections [J]. Pattern Recognition, 2010, 43(6): 1993-2002.

[19] Ding C, Li T, Jordan M I. Nonnegative Matrix Factorization for Combinatorial Optimization: Spectral Clustering, Graph Matching, and Clique Finding [C]//Proc of the 8th IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2008: 183-192.

[20] Arora R, Gupta M R, Kapila A, et al. Similarity-based Clustering by Left-Stochastic Matrix Factorization [J]. Journal of Machine Learning Research, 2013, 14(1): 1715-1746.

Note: Figure translations are in progress. See original paper for figures.