IEICE TRANS. INF. & SYST., VOL.E106–D, NO.1 JANUARY 2023

LETTER

Entropy Regularized Unsupervised Clustering Based on Maximum Correntropy Criterion and Adaptive Neighbors

Xinyu LI[†], **Hui FAN**[†], ***and*** **Jinglei LIU**[††a)], ***Nonmembers***

SUMMARY Constructing accurate similarity graph is an important process in graph-based clustering. However, traditional methods have three drawbacks, such as the inaccuracy of the similarity graph, the vulnerability to noise and outliers, and the need for additional discretization process. In order to eliminate these limitations, an entropy regularized unsupervised clustering based on maximum correntropy criterion and adaptive neighbors (ERMCC) is proposed. 1) Combining information entropy and adaptive neighbors to solve the trivial similarity distributions. And we introduce ℓ_0-norm and spectral embedding to construct similarity graph with sparsity and strong segmentation ability. 2) Reducing the negative impact of non-Gaussian noise by reconstructing the error using correntropy. 3) The prediction label vector is directly obtained by calculating the sparse strongly connected components of the similarity graph Z, which avoids additional discretization process. Experiments are conducted on six typical datasets and the results showed the effectiveness of the method.

key words: *adaptive neighbors, entropy regularized, half-quadratic optimization, maximum correntropy criterion*

1. Introduction

Clustering is an important topic in computer vision and machine learning. With the increase of the amount of data and the complexity of data types, it is more and more time-consuming to label data, so it is necessary to study unsupervised clustering. It can be divided into K-means-based clustering and graph-based clustering [1]. However, it is hard to accurately partition real-world data because it is usually non-linear separable. In order to deal with the complex manifold structure in data, researchers have proposed many graph-based clustering methods [2], [3].

Cai et. al. [4] proposed a regularized graph NMF (GNMF) method to encode geometric information. Pei et al. [5] proposed a concept decomposition method (CFAN) with adaptive neighbors. The basic idea is to integrate an adaptive neighbor regularization constraint into the concept decomposition, and the goal is to extract the representation space that maintains the geometric neighborhood structure of the data. Huang et al. [6] proposed an adaptive graph regularized clustering method based on NMF, which performs matrix decomposition and similarity learning at the same time. By balancing the interaction between the two subtasks in the model, each subtask is iteratively improved based on the results of the other subtask. Wang et al. [7] proposed an unsupervised clustering method that combines information entropy and adaptive neighbors to dynamically learn connected graphs.

Although graph-based clustering have achieved many remarkable achievements, these algorithms still suffer from some shortcomings. Most methods are measured based on distance, and the similarity graph is usually fixed in the subsequent analysis. They lack physical meaning and are very sensitive to the noise contained in the original data. Moreover, the methods mentioned above do not take into account the significant impact of noise and outliers on the clustering results.

2. Related Work

2.1 Adaptive Neighbor

Denote $X = \{x_1, x_2, \ldots, x_n\} \in R^{d\times n}$ as the data matrix and $Y = \{y_1, y_2, \ldots, y_n\} \in R^{d\times n}$ as the clean data matrix. The original data always contain noise and outliers, so we learn similarity graph directly from the clean data Y. Thus, the preliminary neighbor clustering model can be expressed as

$$\min_{z_i^T \mathbf{1}=1, 0\le z_i\le 1, z_{ii}=0} \sum_{j=1}^{n} (\|y_i - y_j\|_2^2 z_{ij}). \tag{1}$$

However, there is a trivial solution to Eq. (1), so that only the point closest to y_i belongs to its neighbor. To solve this problem, we introduce the entropy maximization constraint.

2.2 Information Entropy Maximization

Information entropy is defined as $\Pi(z_i) = \sum_{j=1}^{n}(-z_{ij} \ln z_{ij})$. Since the value of z_{ij} is non-negative, the equation can only reach a minimum when one of the elements has a value of 1 and the others have a value of 0. This sparse overfitting distribution is no different from the trivial solution of the Eq. (1). To avoid this situation, we consider combining the information entropy maximization regularization with the Eq. (1) to reliably and stably fit the current state of similar variables at each step of the optimization process to obtain the optimal similarity graph.

Manuscript received June 23, 2022.
Manuscript revised August 22, 2022.
Manuscript publicized October 6, 2022.

[†]The authors are with the School of Computer Science and Technology, Shandong Technology and Business University, Yantai 264005, P. R. China.
[††]The author is with the School of Computer and Control Engineering, Yantai University, Yantai 264005, P. R. China.
a) E-mail: jinglei_liu@sina.com

DOI: 10.1587/transinf.2022EDL8054

2.3 Spectral Embedding

Denote $F = \{f_1, f_2, \ldots, f_n\} \in R^{n\times c}$ as the indictor matrix, if two data points are similar, their indicator vectors are also closed to each other [8]. It can be expressed as

$$\min \sum_{i,j=1}^{n} \|f_i - f_j\|_2^2 z_{ij}. \tag{2}$$

2.4 Correntropy

Correntropy is a measure of the local and nonlinear similarity of two random variables in information theory [9]. It can be defined as

$$G(P, Q) = E[\kappa(P, Q)] = \int \kappa(p, q) dF_{PQ}(p, q), \tag{3}$$

where κ_σ is a shift-invariant kernel with the bandwidth of σ, $E[\cdot]$ denotes the expectation, and $F_{PQ}(p, q)$ denotes the joint distribution function of (P, Q) [10]. Given a finite number of samples $(p_n, q_n)_{n=1}^N$, the approximate correntropy can be obtained as

$$G_{N,\sigma}(P, Q) = \frac{1}{N} \sum_{n=1}^{N} \Omega\left(p_n - q_n\right), \tag{4}$$

where Gaussian distributed kernel function $\Omega(p - q) = e^{-\frac{(p-q)^2}{2\sigma^2}}$. Correntropy eliminates the negative impact of larger outliers on clustering by adjusting the observation window σ. Correntropy can extract higher-order statistics of data, so as to solve the stability problem of second-order similarity measure [11].

3. Proposed Method

3.1 Model of ERMCC

ERMCC integrates reconstruction error based on correntropy, spectral embedding, adaptive graph construction combining graph regularization and information entropy of similarity matrix into a unified objective function as shown in Eq. (5).

$$\begin{aligned} \max_{Y,F,Z} & \sum_{i=1}^{d} \Omega\left(\sqrt{\sum_{j=1}^{n}\left(X_{ij} - Y_{ij}\right)^2}\right) - \lambda\|f_i - f_j\|_2^2 z_{ij} \\ & - \alpha \sum_{i=1}^{n}\sum_{j=1}^{n}(\|y_i - y_j\|_2^2 z_{ij} - \mu z_{ij} \ln z_{ij}), \\ & s.t.\ \forall i, z_i^T \mathbf{1} = 1, 0 \le z_{ij} \le 1, z_{ii} = 0, \|z_i\|_0 = k. \end{aligned} \tag{5}$$

The advantages of the ERMCC model can be summarized in the following three parts:

(1) Correntropy removes the effects of non-Gaussian noise and outliers by adjusting the kernel bandwidth σ, and it can extract more information from the data for adaptive learning, resulting in more accurate solutions than traditional MSE method.
(2) To solve the problem of noise and outliers and the problem of trivial solutions in the traditional method of constructing similarity graph. The combination of graph regularization based on clean data Y and similarity matrix information entropy reliably and stably fits the current state of similar variables at each step of the optimization process to obtain the optimal similarity graph.
(3) Spectral embedding and ℓ_0-norm constraint are used to satisfy that the constructed sparse similarity graph Z have accurate c connection component. This ensures that the clustering results are obtained while learning the similarity graph, avoiding additional discretization operations.

3.2 Solution of ERMCC

3.2.1 HQ Optimization

Nonlinear and nonconvex problems are difficult to optimize directly, so we use a half-quadratic optimization method for the reconstruction error term [12]. It can be written as

$$\max_{Y,\varrho} \Omega = \sum_{i=1}^{d} \frac{\varrho_i}{\sigma^2}\left(\sum_{j=1}^{n}\left(X_{ij} - Y_{ij}\right)^2 - \varphi\left(\varrho_i\right)\right). \tag{6}$$

When the parameter is fixed, Eq. (6) can be expressed equivalently as

$$\begin{aligned} \min_{Y} \Omega &= -\sum_{i=1}^{d} \varrho_i \sum_{j=1}^{n}\left(X_{ij} - Y_{ij}\right)^2 \\ &= tr(X^T PX - 2Y^T PX + Y^T PY), \end{aligned} \tag{7}$$

where P is a positive diagonal matrix and $P_{ii} = -\varrho_i = e^{-\frac{\sum_{j=1}^{n}\left(X_{ij}-Y_{ij}\right)^2}{2\sigma^2}}$ and $\sigma = \sqrt{\frac{1}{2d}\sum_{i=1}^{d}\sum_{j=1}^{n}\left(X_{ij} - Y_{ij}\right)^2}$. Then Eq. (5) can be transformed as

$$\begin{aligned} \min_{Y,F,Z} & \sum_{i=1}^{d} tr(X^T PX - 2Y^T PX + Y^T PY) + \lambda\|f_i - f_j\|_2^2 z_{ij} \\ & + \alpha \sum_{i=1}^{n}\sum_{j=1}^{n}(\|y_i - y_j\|_2^2 z_{ij} + \mu z_{ij} \ln z_{ij}), \\ & s.t.\ \forall i, z_i^T \mathbf{1} = 1, 0 \le z_{ij} \le 1, z_{ii} = 0, \|z_i\|_0 = k. \end{aligned} \tag{8}$$

3.2.2 Initialize Similarity Matrix Z

Y is initialized to a $d \times n$ random matrix. Introducing the Lagrange multiplier method to Eq. (7), according to the KKT conditions, we can obtain:

$$Y \leftarrow Y\frac{PX}{PY}. \tag{9}$$

Initialize similarity matrix Z, and use the Lagrange mul-

tiplier method to solve Eq. (8). Meanwhile, we denote $u_{ij}^{y} = \|y_i - y_j\|_2^2$, then we can rewrite Eq. (8) as

$$\begin{aligned} &\min_{Z} \alpha \sum_{i=1}^{n}\sum_{j=1}^{n} u_{ij}^{y} z_{ij} + \mu z_{ij} \ln z_{ij} \\ &s.t. \forall i, z_i^T \mathbf{1} = 1, 0 \le z_{ij} \le 1, z_{ii} = 0 \end{aligned} \tag{10}$$

Construct Lagrange function as

$$\mathcal{L} = \alpha \sum_{j=1}^{n} (u_{ij}^{y} z_{ij} + \mu z_{ij} \ln z_{ij}) + \beta(\sum_{j=1}^{n} z_{ij} - 1). \tag{11}$$

Take the partial derivative with respect to z_{ij} and β respectively, then we can obtain:

$$z_{ij} = e^{-\frac{1}{\mu}(u_{ij}^{y} + \frac{\beta}{\alpha}) - 1} = e^{-\frac{\beta}{\alpha\mu} - 1} e^{-\frac{1}{\mu} u_{ij}^{y}}. \tag{12}$$

Substitute Eq. (12) into the partial derivative of $\mathcal{L}$ with respect to β to get

$$e^{-\frac{\beta}{\alpha\mu} - 1} \sum_{m=1}^{n} e^{-\frac{1}{\mu} u_{im}^{y}} = 1 \Leftrightarrow e^{-\frac{\beta}{\alpha\mu} - 1} = \frac{1}{\sum_{m=1}^{n} e^{-\frac{1}{\mu} u_{im}^{y}}}. \tag{13}$$

Substitute Eq. (13) into Eq. (12) to get the initialization solution of z_{ij},

$$\hat{z}_{ij} = \begin{cases} \frac{e^{-\frac{1}{\mu} u_{ij}^{y}}}{\sum_{m=1}^{n} e^{-\frac{1}{\mu} u_{im}^{y}}}, & \text{if } j \neq i; \\ 0 & \text{if } j = i. \end{cases} \tag{14}$$

We set the parameter as $\mu = (\zeta/n)\sqrt{\sum_{i,j=1}^{n} (u_{ij}^{y})^2}$, where ζ is an additional ratio coefficient that adjusts the initial μ. Since $u_{ij}^{y} = \|y_i - y_j\|_2^2$, it is equivalent to setting the value of μ according to different datasets.

3.2.3 Update Variable

Fix Z, Y and update F:

Equation (8) can be transformed as

$$\min_{F} \|f_i - f_j\|_2^2 z_{ij} = \min_{F} \lambda tr(F^T L_Z F). \tag{15}$$

The optimal solution of Eq. (15) is to obtain the orthogonal clustering index matrix F formed by the c eigenvectors of L_Z corresponding to the c minimum eigenvalues.

Fix F, Y and update Z:

Denote $u_{ij}^{f} = \frac{1}{2}\|f_i - f_j\|_2^2$ and $u_{ij} = u_{ij}^{y} + \xi u_{ij}^{f}$, where $\xi = \frac{\lambda}{\alpha}$, then we can rewrite Eq. (8) as

$$\begin{aligned} &\min_{Z} \sum_{i=1}^{n}\sum_{j=1}^{n} \|y_i - y_j\|_2^2 z_{ij} + \mu z_{ij} \ln z_{ij} + \xi \|f_i - f_j\|_2^2 z_{ij} \\ &= \min_{Z} \sum_{i=1}^{n}\sum_{j=1}^{n} u_{ij} z_{ij} + \mu z_{ij} \ln z_{ij} \\ &s.t. \forall i, z_i^T \mathbf{1} = 1, 0 \le z_{ij} \le 1, z_{ii} = 0, \|z_i\|_0 = k. \end{aligned} \tag{16}$$

Referring to the process of initializing the similarity matrix Z, we can obtain

$$\bar{z}_{ij} = \begin{cases} \frac{e^{-\frac{1}{\mu} u_{ij}}}{\sum_{m=1}^{n} e^{-\frac{1}{\mu} u_{im}}}, & \text{if } j \neq i; \\ 0 & \text{if } j = i. \end{cases} \tag{17}$$

The ℓ_0-norm constraint $\|z_i\|_0 = k$ is introduced here to limit the number of non-zero elements in z_i to optimize Eq. (17). On the basis of the obtained $\bar{z}$, elements from k to n are extracted and normalized, while the other n-k elements are discarded, finally we can obtain:

$$z_{ij} = \begin{cases} \frac{e^{-\frac{1}{\mu} u_{ij}}}{\sum_{m=2}^{k+1} e^{-\frac{1}{\mu} u_{im}}}, & \text{if } 2 \le j \le k+1; \\ 0 & \text{if } j = 1 \text{ or } k+1 < j \le n. \end{cases} \tag{18}$$

3.3 Theoretical Analysis of ERMCC

The stopping condition is $rank(L_Z) = n - c$, which means that the obtained similarity graph Z has c connected components. During the iterative process, the value of $rank(L_Z)$ and $n - c$ are re-compared after each update. Setting the initial value of λ to be the same as μ. When $rank(L_Z) < n - c$, it means that the current similarity graph Z does not have enough connected components, and the role of rank constraint should be strengthened, λ should be adjusted to 2λ. Similarly, when $rank(L_Z) > n - c$, λ should be adjusted to $\frac{\lambda}{2}$ to weaken the effect of rank constraints.

4. Experiments

4.1 Experimental Setup

Clustering experiments are conducted on 6 publicly available benchmark datasets, the description of datasets is shown in Table 1. We compare ERMCC with 6 clustering methods: K-means, GNMF [4], CFAN [5], NMFAN [6], ERCAN [7] and ER-F.

4.2 Experimental Result

The comparison of clustering results with various methods on the three evaluation indicators of ACC, NMI and Purity is shown in Table 2. To show the effect of the correntropy, we construct a variant ER-F of ERMCC that excludes the correntropy from the objective function. We used traditional Frobenius norm for the reconstruction error measure and express this term as $\|X - Y\|_F^2$. There are three reasons for

Table 1 Description of datasets.

Type of Dataset	Dataset	Samples	Features	Classes
Face image dataset	JAFFE	676	213	10
	CMU PIE	1166	1024	53
Voice dataset	THCHS30	1440	1024	20
UCI dataset	Balance	625	4	3
	Control	600	60	6
Object image dataset	COIL20	1440	1024	20

Table 2 Experimental results on 6 different datasets.

Dataset	*K*means	GNMF	CFAN	NMFAN	ERCAN	ER-F	ERMCC
				ACC			
JAFFE	76.24	97.65	91.08	76.92	96.71	96.71	**98.12**
CMU PIE	31.55	79.85	46.31	41.54	67.84	68.95	**83.28**
THCHS30	51.10	78.47	60.35	63.77	83.26	85.90	**85.90**
Balance	53.62	51.20	48.48	67.02	62.24	77.92	**78.72**
Control	57.39	58.83	71.00	61.63	68.83	75.50	**75.50**
COIL20	57.55	82.22	65.35	61.82	87.22	87.56	**87.78**
				NMI			
JAFFE	81.48	96.48	89.18	80.65	96.23	96.23	**97.31**
CMU PIE	58.48	92.52	69.39	64.60	84.21	84.49	**93.91**
THCHS30	68.97	85.12	68.03	75.58	90.34	92.58	**92.58**
Balance	14.14	10.53	10.96	31.20	8.02	36.32	**37.49**
Control	65.88	74.21	62.74	69.87	75.80	76.45	**76.45**
COIL20	73.33	89.99	73.09	73.88	94.50	94.50	**94.50**
				Purity			
JAFFE	79.15	97.65	91.08	79.34	96.71	96.71	**98.12**
CMU PIE	37.66	83.79	51.20	46.02	71.10	73.41	**87.22**
THCHS30	55.45	79.65	62.99	67.18	**86.39**	86.04	86.04
Balance	67.62	66.40	64.32	76.62	63.68	80.64	**82.24**
Control	63.51	66.67	71.00	66.78	73.00	75.50	**75.50**
COIL20	61.93	83.96	67.92	66.55	90.00	90.00	**90.00**

the better performance of ERMCC. First, ERMCC eliminates the effect of large differences in results due to the extra discretization process. Then, ERMCC combines information entropy and adaptive neighbors, while introducing ℓ_0-norm constraint and spectral embedding, making the final obtained similarity graph sparse and strongly connected. Finally, compared to ER-F, ERMCC uses the correntropy to address the non-Gaussian noise and outliers which further improves the performance of the method. On some datasets, the performance of ER-F is very close or the same as that of ERMCC, the reason may be that these datasets do not contain non-Gaussian noise.

5. Conclusion

In this letter, an entropy regularized unsupervised clustering based on maximum correntropy criterion and adaptive neighbors (ERMCC) is proposed. It not only removes the extra discretization process and obtains the clustering results directly based on the constructed similarity graph, but also introduces the correntropy to alleviate the influence of non-Gaussian noise and outliers. Experiments have also proved that ERMCC has certain advantages compared with the existing clustering methods.

References

[1] J. Huang, F. Nie, and H. Huang, "Spectral rotation versus k-means in spectral clustering," Proc. AAAI Conference on Artificial Intelligence, vol.27, no.1, pp.431–437, 2013.

[2] Z. Kang, L. Wen, W. Chen, and Z. Xu, "Low-rank kernel learning for graph-based clustering," Knowledge-Based Systems, vol.163, no.1, pp.510–517, 2019.

[3] F. Wang, L. Zhu, C. Liang, J. Li, X. Chang, and K. Lu, "Robust optimal graph clustering," Neurocomputing, vol.378, no.Feb.22, pp.153–165, 2020.

[4] D. Cai, X. He, J. Han, and T.S. Huang, "Graph regularized nonnegative matrix factorization for data representation," IEEE Trans. Pattern Anal. Mach. Intell., vol.33, no.8, pp.1548–1560, 2011.

[5] X. Pei, C. Chen, and W. Gong, "Concept factorization with adaptive neighbors for document clustering," IEEE Trans. Neural Netw. Learn. Syst., vol.29, no.2, pp.343–352, 2018.

[6] S. Huang, Z. Xu, Z. Kang, and Y. Ren, "Regularized nonnegative matrix factorization with adaptive local structure learning," Neurocomputing, vol.382, pp.196–209, 2020.

[7] J. Wang, Z. Ma, F. Nie, and X. Li, "Entropy regularization for unsupervised clustering with adaptive neighbors," Pattern Recognition, vol.125, pp.1–1, 2022.

[8] A. Yuan, M. You, D. He, and X. Li, "Convex non-negative matrix factorization with adaptive graph for unsupervised feature selection," IEEE Trans. Cybern., vol.52, no.6, pp.5522–5534, 2022.

[9] T. Jin, R. Ji, Y. Gao, X. Sun, X. Zhao, and D. Tao, "Correntropy-induced robust low-rank hypergraph," IEEE Trans. Image Process., vol.28, no.6, pp.2755–2769, 2019.

[10] K. Xiong, H.H.C. Iu, and S. Wang, "Kernel correntropy conjugate gradient algorithms based on half-quadratic optimization," IEEE Trans. Cybern., vol.51, no.11, pp.5497–5510, 2021.

[11] W. Liu, P.P. Pokharel, and J.C. Principe, "Correntropy: properties and applications in non-gaussian signal processing," IEEE Trans. Signal Process., vol.55, no.11, pp.5286–5298, 2007.

[12] Y. He, F. Wang, Y. Li, J. Qin, and B. Chen, "Robust matrix completion via maximum correntropy criterion and half-quadratic optimization," IEEE Trans. Signal Process., vol.68, pp.181–195, 2020.