Pairwise constraints cross entropy fuzzy clustering algorithm based on manifold learning and feature selection

Wei Cai¹, Shengbing Xu¹*, LiangJun Zhang², Jiongzhi Liu³, Peixuan Chen¹

¹School of Applied Mathematics, Guangdong University of Technology, guangzhou, china
²Technology Department, Guangdong TipDM Intelligent Technology Co.Ltd, Guangzhou, china
³School of Applied Mathematics, Guangdong University of Technology, guangzhou, china

*Corresponding author’s e-mail: xushengbing@gdut.edu.cn

Abstract. In weakly supervised learning, it is difficult for us to utilize pairwise constraints information in feature selection. In order to solve the problem, we propose Pairwise constraints cross entropy fuzzy clustering algorithm based on manifold learning and feature selection (FCPC-LEFS). There are four phases in our approach: 1) Generate pseudo label; 2) Dimension reduction by Laplacian Eigenmaps; 3) Feature increment and selection; 4) Cross-Entropy semi-Supervised Clustering Based on Pairwise Constraints. We apply our approach to three UCI datasets and a COVID19-CT image dataset. Experiments show that our manifold learning and feature selection method are able to increase improve the clustering performance.

1. Introduction

In many scenarios, it is very expensive for us to label samples. Compared with supervised learning algorithms that require a lot of the labelled samples [1][2][3][4], it’s a wise idea to choose semi-supervised learning algorithms that can make good use of a small few of labeled samples [5].

Semi-supervised fuzzy clustering, which is widely used in computer vision, pattern recognition, speech recognition and all kinds of other fields [6][7][8], is an vital research topic concerned by many scholars [9]. In contrast to unsupervised fuzzy clustering, which is unable to make use of the labeled samples, semi-supervised fuzzy clustering is able to effectually make use of a small few of labeled samples. Generally speaking, there have three types of semi-supervised information that are used in semi-supervised fuzzy clustering: 1) labeled samples; 2) prior membership information; 3) pairwise-constraints information [15].

Feature selection can remove redundant features and improve clustering performance [16]. But Pairwise constraints information cannot be directly used in classic feature selection methods such as Fisher Score like category label information. In real life, however, labels which are category information labeled by the expert are expensive to acquire, but pairwise constraints are easy to calibrate. Therefore, some scholars proposed a feature selection method based on pairwise constraints. [17][18][19]

Based on the above consideration, it is proposed that Pairwise constraints cross entropy fuzzy clustering algorithm based on manifold learning and feature selection in our paper. For the purpose of achieving fast as well as valid fuzzy clustering based on pairwise constraints, we are committed to
improving the salience of features by manifold learning and feature selection. at a word, the primary work in this paper are summarized as follows:

1) We introduce Cross-Entropy semi-Supervised Clustering Based on Pairwise Constraints (CE-sSC).
2) We construct a kind of feature selection method utilizing pairwise constraints.
3) Pairwise constraints cross entropy fuzzy clustering algorithm based on manifold learning and feature selection.
4) We use three datasets from UCI and a COVID19-CT dataset to establish numerical experiments so as to evaluate the performance our method.

2. Related work

2.1. Cross-Entropy semi-Supervised Clustering Based on Pairwise Constraints (CE-sSC)
In semi-supervised fuzzy c-means clustering, pairwise-constrains-samples which belong to the same class or different classes are given in advance. This kind of pairwise constraints [20] are denoted as \( <x_j, x_k> \in ML(Must - Link) \) or \( <x_j, x_k> \in CL(Cannot - Link) \). The mark matrix of pair-wise constraints \( \Gamma = (\Gamma_{jk}) \):

\[
\Gamma_{jk} = \text{sign}(<x_j, x_k>) = \begin{cases} 
-1, & <x_j, x_k> \in CL \\
+1, & <x_j, x_k> \in ML \\
\gamma, & j = k \\
0, & \text{other}
\end{cases}
\]  (1)

Let us first assume that the samples set to be clustered be \( X = \{x_1, x_2, \ldots, x_n\} \) where \( x_j \in \mathbb{R}^d \) \((1 \leq j \leq n)\) in the \( d \)-dimensional Euclidean space. \( c \) is the number of clusters. The objective function of CE-sSC [21] can be expressed as:

\[
J_{CE-sSC}(U,V) = \sum_{i=1}^{c} \sum_{j=1}^{n} \|x_j - v_i\|^2 + \sum_{i=1}^{c} \sum_{j=1}^{n} \Gamma_{ij} \ln u_{ij} 
\]  (2)

Where \( u_{ij} \) is the membership degree, \( v_i \) is the \( i \)-th cluster center, \( x_j \) is the \( j \)-th sample. the membership degree \( U = (u_{ij}) \), cluster center \( V = [v_1,v_2,\ldots,v_c] \), \( 1 \leq i \leq c \), \( 2 \leq c < n \), \( 1 \leq j \leq n \), \( u_{ij} = (\mu_{ij},\cdots,\mu_{ij})^T \) and \( u_{ij} \) must satisfy the constraints:

\[
\sum_{i=1}^{c} u_{ij} = 1, \quad 0 < u_{ij} \leq 1 \quad (3)
\]

By minimizing (1), and using the Lagrange optimization, we may get the following alternative update equations for the cluster center \( v_i \) and the membership degree \( u_{ij} \):

\[
v_i = \frac{\sum_{j=1}^{n} u_{ij} x_j}{\sum_{j=1}^{n} u_{ij}} \quad (4)
\]
2.2. Manifold learning

Laplacian Eigenmaps (LE) [22] is a classical manifold learning method which is applied to dimension reduction. The specific method of Laplacian Eigenmaps is as follows:

$$L = (l_{jk})_{n \times n}$$

(6)

$$L = D - W$$

(7)

Where $L$ is laplacian matrix, $D$ is the degree matrix of a graph, and $W$ is the adjacency matrix of a graph.

$$l_{jk} = \begin{cases} 
\text{deg}(p_j), & j = k \\
-1, & j \neq k \\
0, & \text{other}
\end{cases}$$

(8)

Where $\text{deg}(p_j)$ is the degree of vertex $p_j$.

Let the sample set projected by $X = \{x_1, x_2, \cdots, x_n\}$ be $Z = \{z_1, z_2, \cdots, z_n\}$, and the template of Le be the minimization objective function:

$$E(z) = \sum_{j,k} (z_j - z_k)^2 W_{jk}$$

(9)

Where $W_{jk} = e^{-\frac{||x_j - x_k||^2}{\sigma^2}}$

2.3. Feature selection

Fisher score[23] is a classical feature selection method, the specific method of Fisher score is as follows:

$$FS(X) = \text{tr}\{ (S_w)^{-1} S_b \}$$

(10)

between-class scatter matrix:

$$S_b = \sum_{i=1}^{c} n_i (v_i - v)(v_i - v)^T$$

(11)

within-class scatter matrix:

$$S_w = \sum_{j=1}^{n} n_j (x_j - v)(x_j - v)^T$$

(12)

3. Pairwise constraints cross entropy fuzzy clustering algorithm based on manifold learning and feature selection

3.1. Motivation

Some data have feature redundancy [24], which will have a large impact on clustering performance. It is the traditional method that feature selection is used to solve to feature redundancy. But the
traditional feature selection needs accurate labels rather than pairwise constraints information, how to exclude redundant features by using pairwise constraint information is a topic worthy of discussion.

In order to solve the problem that pairwise constraint information can not be used for feature selection, we propose FCPC-LEFS. The main idea of fuzzy clustering based on pairwise constraints is shown in Figure 1.

![Figure 1. The main ideas of FCPC-LEFS](image)

In Figure 2, influence of manifold learning and feature selection on the performance of pairwise constrained cross entropy clustering algorithm before and after improvement. The stretched image is extracted into 16 features of covid19-ct image data set [25] by hog algorithm. We compare the clustering effect between original pairwise constrained fuzzy clustering and the improved pairwise constrained fuzzy clustering which uses manifold learning and feature selection. The results show that FCPC-LEFS has better clustering effect.

### 3.2. Pairwise constraints cross entropy fuzzy clustering algorithm based on Laplacian Eigenmaps and fisher score (FCPC-LEFS)

Based on the analysis above, we propose pairwise constrained cross entropy fuzzy clustering algorithm based on Laplacian Eigenmaps and fisher score.

- The first is to generate pseudo tags.
Get the predicted category tag $\hat{y}$ by Fuzzy c-means clustering, using $\hat{y}$. Generating the predicted mark matrix of pair-wise constraints $\hat{\Gamma}$.

$$\hat{\Gamma} - \hat{\Gamma} = \hat{\Gamma}^{\Theta}$$  \hspace{1cm} (13)

Search the row or column index of 0 elements in the matrix $\hat{\Gamma}^{\Theta}$ to find their corresponding index. These are the labels that can be used for feature selection, that is, the effective category labels.

- Second, manifold learning.
- Dimension reduction of original data $X$ to get $Z$.
- Then, Feature increment and feature selection.

The dimensionality-reduced data and the original data are combined to form new data. For convenience, the new data is still recorded as $X$:

$$\hat{X} = [X \ Z]$$  \hspace{1cm} (14)

The saliency of each feature is calculated by fisher score using effective labels $\hat{y}$ and corresponding samples in new data. The feature with lower 50% Fisher score value is discarded from all features.

- Semi-supervised Clustering Based on Pairwise Constraints using Cross-Entropy semi-Supervised Clustering Based on Pairwise Constraints get $U$ and $V$.

3.3. Algorithm flow (FCPC-LEFS)

1) Generate pseudo label by (13).
2) manifold learning: Dimension reduction by Laplacian Eigenmaps (9).
3) Feature increment: Add dimension reduction data to original data to form new data by (10).
4) Feature selection: The score of each feature of the new data is calculated by discarding the insignificant features through the Fisher score (13).
5) Given suppression coefficient $\gamma$, the number of cluster centers $c$, error-threshold $E$, weighted power exponent $m$, maximum number of iterations $T$, and currently iterative order $t = 1$;
6) initialize: the cluster center $\nu_i$, the membership matrix $u_{ij}$, The mark matrix of pair-wise constraints $\Gamma$;
7) Update $\nu_i$ on fixing $u_{ij}$ by (4);
8) Update $u_{ij}$ by (5);
9) Calculate the objective function value by (2). $e = J(t) - J(t-1)$ when $t > 1$. If $e < E$ or $t = T$ algorithm ends; otherwise $t = t + 1$ then go back to step 7.

4. Experiment

4.1. Data set
UCI data set (iris and wine).

4.2. Evaluation Measures
F-measure is the weighted harmonic average of precision and recall, which is a common evaluation standard in the field of IR (information retrieval). When the value of F-measure is higher, the corresponding clustering method is more effective. Therefore, We used $F_1$-measure[26] to evaluate the effectiveness of the algorithm.
4.3. Algorithm Performance

![Algorithm Performance](image)

Figure 3. Algorithm effect comparison.

The result in Figure 3 shows that pairwise constrained cross entropy fuzzy clustering algorithm based on Laplacian Eigenmaps and fisher score has the better performance.

5. Conclusion

Selecting valid or significant features can avoid over-fitting, which is an important role in improving the robustness and interpretability of the algorithm. This paper proposes pairwise constraints fuzzy clustering algorithm based on manifold learning and feature selection, which can get better pairwise-constrained fuzzy clustering results.

Acknowledgment

This work was supported by the Innovation Training Project of Guangdong University of Technology under Grant xj202011845507 and D202012040955485708. In addition, it is supported by the Scientific and Technological Innovation Strategy Special Fund Project of Guangdong Province under Grant pdjh2020b0182.

References

[1] Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms[C]//Proceedings of the 23rd international conference on Machine learning. 2006: 161-168.
[2] Weinberger K Q. Distance Metric Learning for Large Margin Nearest Neighbor Classification[J]. Jmlr, 2009, 10.
[3] Suykens J A K, Vandewalle J. Least squares support vector machine classifiers[J]. Neural processing letters, 1999, 9(3): 293-300.
[4] Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in neural information processing systems, 2012, 25(2).
[5] Zhu X, Goldberg A B. Introduction to semi-supervised learning[J]. Synthesis lectures on artificial intelligence and machine learning, 2009, 3(1): 1-130.
[6] Mostafa J, Lam W. Automatic classification using supervised learning in a medical document filtering application[J]. Information Processing & Management, 2000, 36(3): 415-444.
[7] Balcan M F, Blum A, Choi P P, et al. Person identification in webcam images: An application of semi-supervised learning[C]//ICML 2005 Workshop on Learning with Partially Classified Training Data. 2005, 2: 6.
[8] Lin B. Online semi-supervised learning in contextual bandits with episodic reward[J]. arXiv preprint arXiv:2009.08457, 2020.
[9] Bezdek J C, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm[J]. Computers & Geosciences, 1984, 10(2-3):191-203.

[10] FAN Jiulun. A brief overview on suppressed fuzzy C-means clustering[J]. JOURNAL OF XI’AN UNIVERSITY OF POSTS AND TELECOMMUNICATIONS, 2014, 019(003):1-5.

[11] Grira N, Crucianu M, Boujemaa N. Semi-Supervised Fuzzy Clustering with Pairwise-Constrained Competitive Agglomeration[C]// The 14th IEEE International Conference on Fuzzy Systems, 2005. FUZZ ’05. IEEE, 2005.

[12] Yasunori E, Yukihiro H, Makito Y, et al. On semi-supervised fuzzy c-means clustering.[C]// IEEE International Conference on Fuzzy Systems. IEEE, 2009.

[13] Bora D J, Gupta A K. A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm[J]. International Journal of Emerging Trends & Technology in Computer ence, 2014, 10(2):108-113.

[14] Pawan K, Deepika S. Comparative analysis of FCM and HCM algorithm on Iris data set[J]. International Journal of Computer Applications, 2010, 5(2):33-37.

[15] Śmieja M, Struski Ł, Figueiredo M A T. A classification-based approach to semi-supervised fuzzy clustering with pairwise constraints[J]. Neural Networks, 2020.

[16] Cai W, Xu S, Liu J, et al. An adaptive approach of Feature Selection applied to Semi-Supervised Fuzzy Clustering[C]//Proceedings of the 2020 4th International Conference on Electronic Information Technology and Computer Engineering. 2020: 723-727.

[17] Noorie Z, Afsari F. Sparse feature selection: Relevance, redundancy and locality structure preserving guided by pairwise constraints[J]. Applied Soft Computing, 2020.

[18] Samah H, Mariam K, Denis H, et al. Feature selection approach based on hypothesis-margin and pairwise constraints[J]. 2018 IEEE Middle East and North Africa Communications Conference (MENACOMM),2018.

[19] Sheikhpour R, Sarram M A, Gharaghani S, et al. Constraint score for semi-supervised feature selection in ligand and receptor-based QSAR on serine/threonine-protein kinase PLK3 inhibitors[J]. Chemometrics and Intelligent Laboratory Systems, 2017: 31-40.

[20] N. Grira, M. Crucianu and N. Boujemaa, "Semi-Supervised Fuzzy Clustering with Pairwise-Constrained Competitive Agglomeration," The 14th IEEE International Conference on Fuzzy Systems, 2005. FUZZ ’05., Reno, NV, USA, 2005, pp. 867-872, doi: 10.1109/FUZZY.2005.1452508.

[21] Li C M, Xu S B, Hao Z F. Cross-Entropy semi-supervised clustering based on pairwise constraints [j][J]. Pattern Recognition and Artificial Intelligence, 2017, 30(7): 598-608.

[22] Belkin M, Niyogi P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation[J]. 2003.

[23] Gu Q, Li Z, Han J. Generalized Fisher Score for Feature Selection[C]// Conference on Uai. AUAI Press, 2012.

[24] Pal N R, Bezdek J C. On cluster validity for the fuzzy c-means model[J]. IEEE Transactions on Fuzzy Systems, 2002, 3(3):370-379.

[25] Zhao J, Zhang Y, He X, et al. Covid-ct-dataset: a ct scan dataset about covid-19[J]. arXiv preprint arXiv:2003.13865, 2020.

[26] Larsen B, Aone C. Fast and effective text mining using linear-time document clustering[C]//Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. 1999: 16-22.