Engineering Applications of Artificial Intelligence 81 (2019) 107–116

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Confidence-weighted safe semi-supervised clustering ☆

Haitao Gan [a,*], Yingle Fan [a], Zhizeng Luo [a], Rui Huang [b], Zhi Yang [c]

[a] School of Automation, Hangzhou Dianzi University, Hangzhou, China
[b] School of Science and Engineering, Chinese University of Hong Kong, Shenzhen, Shenzhen, China
[c] School of Computer Science, Hubei University of Technology, Wuhan 430068, China

ARTICLE INFO

Keywords:
Semi-supervised clustering
Safe mechanism
Confidence weight
Normalized confusion matrix

ABSTRACT

In this paper, we propose confidence-weighted safe semi-supervised clustering where prior knowledge is given in the form of class labels. In some applications, some samples may be wrongly labeled by the users. Therefore, our basic idea is that different samples should have different impacts or confidences on the clustering performance. In our algorithm, we firstly use unsupervised clustering to perform the dataset partition and compute the normalized confusion matrix Nc. Nc is used to estimate the safe confidence of each labeled sample based on the assumption that a correctly clustered sample should have a high confidence. Then we construct a local graph to model the relationship between the labeled and its nearest unlabeled samples through the clustering results. Finally, a confidence-weighted fidelity term and a graph-based regularization term are incorporated into the objective function of unsupervised clustering. In this case, on the one hand, the outputs of the labeled samples with high confidences are restricted to be the given prior labels. On the other hand, the outputs of the labeled ones with low confidences are forced to approach those of the local homogeneous unlabeled neighbors modeled by the local graph. Hence, the labeled samples are expected to be safely exploited which is the goal of safe semi-supervised clustering. To verify the effectiveness of our algorithm, we carry out some experiments over several datasets by comparison to the unsupervised and semi-supervised clustering methods and achieve the promising results.

1. Introduction

Recently, safe semi-supervised learning (S3L) has attracted much attention in machine learning field. In some scenarios, the traditional semi-supervised learning (SSL) methods may perform worse than the corresponding supervised learning (SL) methods which restricts the practical applications of SSL. In other words, unlabeled samples may be harmful to the performance. Therefore, S3L tries to develop different safe mechanisms to realize that the learning performance is never inferior to that of SL by safely exploiting the unlabeled samples. Due to the merit, S3L extends the application scopes of SSL. In fact, some previous studies (Gan et al., 2013a; Cohen et al., 2004; Singh et al., 2009; Yang and Priebe, 2011) have analyzed the negative impact of unlabeled samples on the learning performance in theoretical and empirical aspects. For the safe exploitation of the unlabeled samples, Li and Zhou proposed two S3L methods in 2011, named S3VM_*us* (Li and Zhou, 2011a) and safe semi-supervised SVMs (S4VMs) (Li and Zhou, 2011b). S3VM_*us* introduced a safe mechanism by selecting the helpful unlabeled samples based on a hierarchical clustering method. Different from S3VM_*us* which only found one optimal low-density separator, S4VMs constructed multiple S3VM candidates simultaneously to reduce the risk of the unlabeled samples. S3VM_*us* and S4VMs both yielded the promising results and reached the goal of S3L. Up to now, several S3L methods (Wang and Chen, 2013; Li et al., 2016b; Wang et al., 2016; Li et al., 2016a; Dong et al., 2016; Li et al., 2017) have been proposed to alleviate the harm of the unlabeled samples for SSL. However, these S3L methods were mainly designed for semi-supervised classification. Furthermore, Li et al. (2017) proposed safe semi-supervised regression (SAFER) which was used for semi-supervised regression. That is to say, the past studies mainly focused on classification and regression. Specifically, there is not related work for semi-supervised clustering.

In fact, past decades have witnessed the successfulness of semi-supervised clustering in the various practical applications. The goal of semi-supervised clustering is to fully utilize the prior knowledge to aid the clustering procedure, such as class labels and pair-wise constraints. A lot of semi-supervised clustering methods (Gan et al., 2015; Zhang

☆ One or more of the authors of this paper have disclosed potential or pertinent conflicts of interest, which may include receipt of payment, either direct or indirect, institutional support, or association with an entity in the biomedical field which may be perceived to have potential conflict of interest with this work. For full disclosure statements refer to https://doi.org/10.1016/j.engappai.2019.02.007.

* Corresponding author.
E-mail addresses: htgan@hdu.edu.cn (H. Gan), fan@hdu.edu.cn (Y. Fan), luo@hdu.edu.cn (Z. Luo), ruihuang@cuhk.edu.cn (R. Huang), D201077542@hust.edu.cn (Z. Yang).

https://doi.org/10.1016/j.engappai.2019.02.007
Received 12 November 2017; Received in revised form 30 December 2018; Accepted 6 February 2019
Available online xxxx
0952-1976/© 2019 Elsevier Ltd. All rights reserved.

and Lu, 2009; Basu et al., 2002; Chen and Feng, 2012; Givoni and Frey, 2009; Bensaid et al., 1996; Pedrycz and Waletzky, 1997) are developed based on the traditional unsupervised clustering methods, such as *k*-means (Hartigan and Wong, 1979), Gaussian mixture models (GMM) (Chen et al., 2011), Fuzzy *c*-Means (FCM) (Bezdek, 1981). Traditional semi-supervised clustering generally gives the hypothesis that the prior knowledge is benefit to the clustering performance. However, the collected prior knowledge (e.g., wrongly labeled samples and noise) may result in the performance degeneration as mentioned in the semi-supervised classification and regression. Yin et al. (2010) have discussed the negative impact of noisy pair-wise constraints and pointed out that the wrong prior knowledge would yield the inferior clustering performance.

Based on the mentioned-above two aspects, it is meaningful and worthy to design a safe semi-supervised clustering method which can outperform the corresponding unsupervised and semi-supervised clustering methods. Recently, Gan et al. (2018) developed Local Homogeneous Consistent Safe Semi-Supervised FCM (LHC-S^{3}FCM) where the class labels are given as the prior knowledge. A new graph-based regularization term was built for LHC-S^{3}FCM which meant that the outputs of the labeled sample and its nearest homogeneous unlabeled ones should be similar. However, it is implied that the labeled samples equally hurt the clustering performance in LHC-S^{3}FCM.

Hence, we invent confidence-weighted safe semi-supervised clustering in this paper. Different from LHC-S^{3}FCM, our basic idea is that different samples should have different impacts or confidences on the performance degeneration. In our algorithm, we firstly use unsupervised clustering to perform the dataset partition and compute the normalized confusion matrix based on the clustering results. The probability distribution in the normalized confusion matrix is used to compute the safe confidence of each labeled sample based on the assumption that a correctly clustered sample should have a high confidence. Then we construct a local graph which is similar to the literature (Gan et al., 2018). The graph can be used to model the relationship between the labeled and its nearest unlabeled samples through the clustering results. Finally, a confidence-weighted fidelity term and a graph-based regularization term are incorporated into the objective function of unsupervised clustering. On the one hand, the outputs of the labeled samples with high confidences are restricted to be the given prior labels. On the other hand, the outputs of the labeled samples with low confidences are forced to approach that of the local homogeneous unlabeled neighbors modeled by the local graph. In this sense, the outputs of the labeled samples in our algorithm are a tradeoff between the given labels and the outputs of local nearest neighbors. Hence, the labeled samples are expected to be safely exploited which is the goal of safe semi-supervised clustering. The main contributions of the paper can be summarized as:

1. We develop a confidence-weighted safe semi-supervised clustering method which can safely exploit the labeled samples.
2. The safe confidences of the labeled samples are estimated by unsupervised clustering which is free from the wrong or noise labels.
3. A local graph is constructed to model the relationship between the labeled and its nearest unlabeled samples and the graph structure is used to safely exploit the risky prior knowledge.

The structure of the paper is organized as: Section 2 will review the related work. Then we will present the details of our algorithm in Section 3. Section 4 will report the results on several databases. Finally, Section 5 will give the conclusion and future work.

2. Related work

In S3L, since Li and Zhou proposed S3VM_*us* (Li and Zhou, 2011a) and S4VMs (Li and Zhou, 2011b) in 2011, several S3L methods have been developed for safely exploiting the unlabeled samples and achieved the promising results. Wang and Chen (2013) developed a safety-aware SSCCM (SA-SSCCM) which is extended from the semi-supervised classification method based on class membership (SSCCM). The performance of SA-SSCCM is never significantly inferior to that of SL and seldom significantly inferior to that of SSL. Li et al. (2016a) proposed safe semi-supervised learning under different multivariate performance measures, such as Top-*k* precision, AUC, F1 score. Meanwhile, Li et al. (2016b) builded a large margin approach named large margin graph quality judgment (LEAD). LEAD constructed multiple graphs simultaneously and tried to exploit the graphs with large margin while keep the graphs with small margin to be rarely exploited. Gan et al. (2016) proposed risk-based safe Laplacian regularized least squares (RsLapRLS) which tried to assign the different risk degrees to different unlabeled samples. Recently, Li et al. (2017) proposed safe semi-supervised regression (SAFER) which was designed for semi-supervised regression. SAFER attempted to learn a safe prediction from multiple semi-supervised regressors and yielded the desired performance.

In semi-supervised clustering, the proposed methods can generally be divided into the following categories: (1) distance-based approach; (2) constraint-based approach; (3) hybrid-based approach.

The distance-based approach studies how to learn a distance measure which should satisfy the given prior knowledge. Many researchers proposed the distance learning methods for semi-supervised clustering in the past years (Yin et al., 2010; Yan et al., 2012; Yin et al., 2012; de Amorim and Mirkin, 2012; Ding et al., 2014). Yin et al. (2010) proposed an adaptive semi-supervised clustering kernel method based on metric learning(SCKMM) which utilized the pair-wise constraints. de Amorim and Mirkin (2012) developed Minkowski metric weighted *k*-means which used the Minkowski metric to measure the distance between two samples. The Minkowski metric was learned from the samples in a semi-supervise manner. Yan et al. (2012) invented a novel search-based semi-supervised clustering method which learned the multi-viewpoint based similarity measure. Ding et al. (2014) utilized the pair-wise constraints to construct an adaptive similarity matrix and developed a semi-supervised spectral clustering method. Kalintha et al. (2017) proposed kernelized evolutionary distance metric learning (K-EDML) which learned a kernelized distance metric and obtained the promising clustering results on the non-linear separable datasets.

The constraint-based approach studies how to revise the objective function or initialize the cluster centers to guide the clustering process. Basu et al. (2002) developed a semi-supervised version of *k*-means called seeded-*k*means, which used the labeled seeds to compute the initial cluster centers. Pedrycz and Waletzky (1997) proposed semi-supervised FCM (SSFCM) with a fidelity term Based on the FCM algorithm. SSFCM implemented a tradeoff between the unsupervised outputs and given prior labels. Mai and Ngo (2015) presented another semi-supervised version of FCM, named SFCM. Different from SSFCM, SFCM introduced a fidelity term which moved the cluster centers to the predefined ones obtained by the labeled samples. Mai and Ngo (2018) proposed Semi-supervised Kernel FCM in Feature space (SKFCM-F) where the rudimentary centers were estimated through the labeled samples. Martinez-Uso et al. (2010) presented semi-supervised GMM (semiGMM) which employed the labeled samples to initialize the parameters of multiple Gaussian components. The results on image segmentation showed the effectiveness of semiGMM by considering the manually labeled pixels. Gan et al. (2015) introduced a semi-supervised locally consistent Gaussian mixture models (Semi-LCGMM) and also applied the method to image segmentation. Jia et al. (2018) proposed a semi-supervised spectral clustering which used the labeled samples to construct a block-diagonal matrix.

Furthermore, some researchers studied the hybrid-based approach (Basu et al., 2004; Bilenko et al., 2004; Zhang et al., 2015; Wei et al., 2017). Basu et al. (2004) proposed a hidden Markov random fields-based probabilistic framework for semi-supervised clustering which combined the distance-based and constraint-based approaches. In this method, Bregman divergence could be learned for the distance measure. Zhang et al. (2015) proposed graph-based GMM which utilized

the labeled samples to learn a Mahalanobis distance and initialize Gaussian distribution parameters. Wei et al. (2017) developed a semi-supervised clustering ensemble approach which taken both pairwise constraints and distance measure into account. Image pixels clustering results verified the effectiveness of the proposed hybrid approach.

3. Confidence-weighted safe SSFCM (CS3FCM)

3.1. Motivation

Traditional semi-supervised clustering generally assumes that the labeled samples are always benefit to the performance improvement. However, in some scenarios, the samples may be wrongly labeled by the users. In other words, the sample labels may be different from the true ones. Traditional semi-supervised clustering does not consider the risk of the wrongly labeled samples. And it is a reasonable assumption that different samples should have the different impacts or safe confidences on the performance.

Formally, given a labeled subset $X_l = [x_1, \ldots, x_l]$ and unlabeled subset $X_u = [x_{l+1}, \ldots, x_n]$, each labeled sample x_k will have a label $y_k \in \{1, \ldots, c\}$ with the number of classes c. When the safe confidence of x_k is high, x_k may be helpful and the corresponding output should be approach to the given label y_k. In this case, the confidence weight of x_k should be large. Otherwise, x_k may be risky and the corresponding output should be approach to that of FCM. In this case, the confidence weight of x_k should be small. Since the outputs obtained by FCM are inconsistent with the given labels, the outputs of the labeled sample in our algorithm cannot be forced to approach that in FCM directly. Our idea is that we construct a local graph to model the relationship between the labeled and its nearest homogeneous unlabeled samples which belong to the same cluster in FCM. And we restrict the outputs of the risky labeled samples to be that of the homogeneous unlabeled ones through the modeled relationship. Hence the risk of the labeled samples is expected to be reduced and one can safely use the prior knowledge for semi-supervised clustering.

3.2. Confidence estimation

Based on the above analysis, we propose Confidence-weighted Safe SSFCM (CS3FCM). Firstly, we employ unsupervised clustering (i.e., FCM) to partition the whole dataset into c clusters and obtain the partition matrix $\widetilde{U} = [\widetilde{u}]_{c\times n}$ and the predicted cluster labels $\widetilde{Y} = [\widetilde{y}_1, \ldots, \widetilde{y}_l, \widetilde{y}_{l+1}, \ldots, \widetilde{y}_n]$. According to the results, we can compute the normalized confusion matrix Nc between the give labels $y_k|_{k=1}^{l}$ and predictions $\widetilde{y}_k|_{k=1}^{l}$. Here we employ the Kuhn–Munkres algorithm (Lovasz and Plummer, 1986) to map the obtained labels $\widetilde{y}_k|_{k=1}^{l}$ to the equivalent labels $\widehat{y}_k|_{k=1}^{l}$ which is consistent with the given labels $y_k|_{k=1}^{l}$. The map function is denoted as $\varphi(\widetilde{y}_k) = \widehat{y}_k$. Nc can be represented as

$$Nc = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1c} \\ p_{21} & p_{22} & \cdots & p_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ p_{c1} & p_{c2} & \cdots & p_{cc} \end{bmatrix} \tag{1}$$

where $\sum_{j=1}^{c} p_{ij} = 1$ and $0 \le p_{ij} \le 1$. p_{ij} denotes the percentage in the ith class that be clustered into the jth class.

For a labeled sample x_k, it may have a high safe confidence if $y_k = \widehat{y}_k$ and $p_{y_k,\widehat{y}_k}$ is large. Moreover, if the fuzzy degree $\widetilde{u}_{\widetilde{y}_k,k}$ which denotes the probability of x_k belonging to the $\widetilde{y}_k$th cluster is large, x_k will be helpful. In this case, x_k is beneficial to improve the clustering performance. Otherwise, x_k may be risky. Hence, we define the weight s_k of x_k as:

$$s_k = \begin{cases} p_{y_k,\widehat{y}_k} \times \widetilde{u}_{\widetilde{y}_k,k} & \text{if } y_k = \widehat{y}_k \\ p_{y_k,\widehat{y}_k} \times (1 - \widetilde{u}_{\widetilde{y}_k,k}) & \text{otherwise} \end{cases} \tag{2}$$

3.3. Graph construction

Since a local graph is used to model the relationship between the labeled and unlabeled samples, we need find the homogeneous nearest unlabeled neighbors for the labeled samples according to the Euclidean distance and the clustering results of FCM. Then the edge weight of the local graph $W = [w_{kr}]_{n\times n}$ can be calculated as:

$$w_{kr} = \begin{cases} \exp\{-\dfrac{\|x_k - x_r\|_2^2}{\sigma^2}\} & \text{if } x_r \in N_p(x_k) \text{ and } \widetilde{y}_k = \widetilde{y}_r \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $N_p(x_k)$ denotes the data sets of p nearest neighbors of x_k. x_k and x_r respectively represent the labeled and unlabeled samples.

3.4. Objective function

We firstly give the objective function of SSFCM:

$$J = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^m d_{ik}^2 + \alpha \sum_{k=1}^{n}\sum_{i=1}^{c} (u_{ik} - f_{ik}b_k)^2 d_{ik}^2 \tag{4}$$

where m is the fuzzy degree with $m > 1$. $U = [u_{ik}] \in R^{c\times n}$ is the partition matrix and $d_{ik} = \|x_k - v_i\|_2$ denotes the distance between the kth sample x_k and the ith cluster center v_i. α is a tradeoff parameter which controls a balance between the unsupervised component and given labels. $B = [b_k]_{1\times n}$ is a label indicator where $b_k = 1$ if x_k is labeled and $b_k = 0$, otherwise. $F = [f_{ik}]_{c\times n}$ denotes the fuzzy degrees of the labeled samples where $f_{ik} = 1$ if $y_k = i$ and $f_{ik} = 0$ otherwise.

Based on the analysis in Section 3.1, we can formulate the objective function of our algorithm as follows:

$$\begin{aligned} J_c = &\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^m d_{ik}^2 + \lambda_1 \sum_{k=1}^{l} s_k \sum_{i=1}^{c} (u_{ik} - f_{ik})^2 d_{ik}^2 \\ &+ \lambda_2 \sum_{k=1}^{l} \frac{1}{s_k} \sum_{r=l+1}^{n} w_{kr} \sum_{i=1}^{c} (u_{ik} - u_{ir})^2 \end{aligned} \tag{5}$$

$$\text{Subject to:} \quad \sum_{i=1}^{c} u_{ik} = 1, \forall k = 1, \ldots, n$$

where λ_1 and λ_2 are the regularization parameters.

As can be seen from Eq. (5), the second term forces the output u_{ik} of a labeled sample x_k to be the given label y_k if the sample x_k has a large weight s_k. Otherwise, the corresponding output u_{ik} will approach to that of the local homogeneous unlabeled samples indicated by the third term.

3.5. Solution

In this paper, we set the fuzzy degree m to 2 for simplifying the optimization problem (5) as in Pedrycz and Waletzky (1997) and the simplified optimization problem can be resolved through the alternating iterative method. Certainly, one can resolve the problem with $m \neq 2$ by some optimization algorithm, such as genetic algorithm.

(1) The solution of u_{ik}

Firstly, the solution of u_{ik} can be obtained through the Lagrangian multiplier method. we can give the Lagrangian multiplier function as:

$$\begin{aligned} \mathcal{L} = &\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^2 d_{ik}^2 + \lambda_1 \sum_{k=1}^{l} s_k \sum_{i=1}^{c} (u_{ik} - f_{ik})^2 d_{ik}^2 \\ &+ \lambda_2 \sum_{k=1}^{l} \frac{1}{s_k} \sum_{r=l+1}^{n} w_{kr} \sum_{i=1}^{c} (u_{ik} - u_{ir})^2 - \gamma(\sum_{i=1}^{c} u_{ik} - 1) \end{aligned} \tag{6}$$

For the labeled sample x_k, the relevant part can be given as:

$$\begin{aligned} \mathcal{L}_1 = &\sum_{k=1}^{l}\sum_{i=1}^{c} u_{ik}^2 d_{ik}^2 + \lambda_1 \sum_{k=1}^{l} s_k \sum_{i=1}^{c} (u_{ik} - f_{ik})^2 d_{ik}^2 \\ &+ \lambda_2 \sum_{k=1}^{l} \frac{1}{s_k} \sum_{r=l+1}^{n} w_{kr} \sum_{i=1}^{c} (u_{ik} - u_{ir})^2 - \gamma(\sum_{i=1}^{c} u_{ik} - 1) \end{aligned} \tag{7}$$

By taking the derivative of $\mathcal{L}_1$ with respect to u_{ik} and setting it to 0, we have the following equation:

$$2u_{ik}d_{ik}^2 + 2\lambda_1 s_k(u_{ik} - f_{ik})d_{ik}^2 + 2\frac{\lambda_2}{s_k}\sum_{r=l+1}^{n} w_{kr}(u_{ik} - u_{ir}) - \gamma = 0 \tag{8}$$

Therefore, the solution of u_{ik} for the labeled sample x_k can be obtained as:

$$u_{ik} = \frac{p_{ik} + \frac{1-\sum_{i=1}^{c}\frac{p_{ik}}{q_{ik}}}{\sum_{i=1}^{c}\frac{1}{q_{ik}}}}{q_{ik}} \tag{9}$$

where $p_{ik} = \lambda_1 s_k f_{ik} d_{ik}^2 + \frac{\lambda_2}{s_k}\sum_{r=l+1}^{n} w_{kr}u_{ir}$ and $q_{ik} = d_{ik}^2 + \lambda_1 s_k d_{ik}^2 + \frac{\lambda_2}{s_k}\sum_{r=l+1}^{n} w_{kr}$.

For the unlabeled sample x_r, the relevant part can be written as:

$$\mathcal{L}_2 = \sum_{r=l+1}^{n}\sum_{i=1}^{c} u_{ir}^2 d_{ir}^2 + \lambda_2 \sum_{k=1}^{l}\frac{1}{s_k}\sum_{r=l+1}^{n} w_{kr}\sum_{i=1}^{c}(u_{ik} - u_{ir})^2 - \gamma\left(\sum_{i=1}^{c} u_{ik} - 1\right) \tag{10}$$

By setting the derivative of $\mathcal{L}_2$ with respect to u_{ir} to 0, we can achieve the following equation:

$$2u_{ir}d_{ir}^2 - 2\lambda_2 \sum_{k=1}^{l}\frac{w_{kr}}{s_k}(u_{ik} - u_{ir}) - \gamma = 0 \tag{11}$$

The solution of u_{ir} for the unlabeled sample x_r can be given as:

$$u_{ir} = \frac{z_{ir} + \frac{1-\sum_{i=1}^{c}\frac{z_{ir}}{t_{ir}}}{\sum_{i=1}^{c}\frac{1}{t_{ir}}}}{t_{ir}} \tag{12}$$

where $z_{ir} = \lambda_2 \sum_{k=1}^{l}\frac{w_{kr}}{s_k}u_{ik}$ and $t_{ir} = d_{ir}^2 + \lambda_2\sum_{k=1}^{l}\frac{w_{kr}}{s_k}$.

(2) The solution of v_i

By taking the derivative of J_c with respect to v_i based on the distance measure $d_{ik}^2 = \|x_k - v_i\|_2^2$, we can have the following equation:

$$\frac{\partial J_c}{\partial v_i} = -2\sum_{k=1}^{n} u_{ik}^2(x_k - v_i) - 2\lambda_1 \sum_{k=1}^{l} s_k(u_{ik} - f_{ik})^2(x_k - v_i) \tag{13}$$

By setting the derivative to 0, we can yield the following solution:

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^2 x_k + \lambda_1 \sum_{k=1}^{l} s_k(u_{ik} - f_{ik})^2 x_k}{\sum_{k=1}^{n} u_{ik}^2 + \lambda_1 \sum_{k=1}^{l} s_k(u_{ik} - f_{ik})^2} \tag{14}$$

Hence, we can obtain the optimal partition matrix U and cluster centers V by iteratively computing u_{ik} and v_i. The iteration process will terminate when $|J_c^{(t)} - J_c^{(t-1)}| < \eta$ or the maximum number of iterations $Maxiter$ is reached, where η is a pre-defined threshold. Fig. 1 gives a plot of the convergence process on Iris dataset. As shown in the figure, our algorithm will converge after several iterations (i.e., $t = 66$) and it verifies the rationality of the convergence criterion. The iterative process of our algorithm can be summarized as in Algorithm 1.

4. Experimental analysis

In this section, we carry out a series of experiments over several datasets to evaluate the performance of our algorithm. The performance is measured through the clustering accuracy and used to verified the effectiveness of our algorithm by comparison to the following methods:

- k-means (Jain, 2010)
- GMM (Bouman, 1997)
- FCM (Bezdek, 1981)
- seeded-$k$means (Basu et al., 2002)
- semiGMM (Martinez-Uso et al., 2010)
- SSFCM (Pedrycz and Waletzky, 1997)

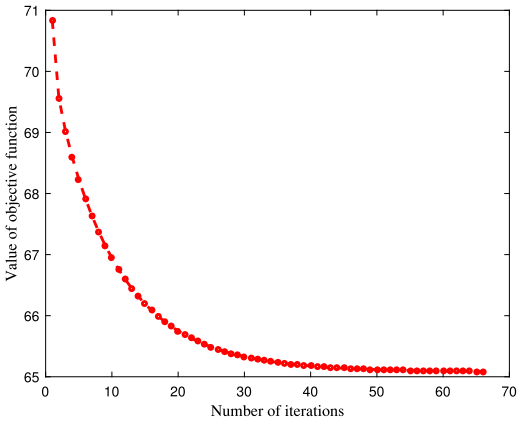

Fig. 1. A convergence illustration of our algorithm on Iris dataset.

Algorithm 1 CS3FCM

Input: A labeled subset $X_l = [x_1, \cdots, x_l]$ with the corresponding labels $Y_l = [y_1, \cdots, y_l]^T$ and unlabeled subset $X_u = [x_{l+1}, \cdots, x_n]$. The parameters λ_1, λ_2, p, σ, η, and $Maxiter$.

Output: Optimal partition matrix U.

```
Perform FCM on the dataset X_l ∪ X_u to yield the cluster result Ỹ, partition matrix Ũ and normalized confusion matrix Nc;
Compute the weight s_k;
Construct the local graph W;
Initialize the cluster center V^(0) by computing the mean of the labeled samples in each class;
for t = 1 : Maxiter do
    Update u_ik^(t) using Eq. (9) and Eq. (12);
    Update v_i^(t) using Eq. (14);
    Compute the value of J_c^(t) using Eq. (5);
    if |J_c^(t) − J_c^(t−1)| < η then
        return U.
    end if
end for
```

- SFCM (Mai and Ngo, 2015)
- SKFCM-F (Mai and Ngo, 2018)
- LHC-S^{3}FCM (Gan et al., 2018)

The used datasets include two artificial ones (i.e., Gauss50 and Gauss50x) (Gan et al., 2013b) and fourteen real ones (i.e., UCI Frank and Asuncion, 2010 and USPS). The statistical details are given in Table 1. For each dataset, we randomly select 20% to form the labeled subset and the rest to form the unlabeled subset. In order to analyze the harm of the wrongly labeled samples, some labeled samples are randomly labeled with wrong labels which are different from the true labels. The ratio of the wrongly labeled samples changes from 0%–30% with step size 5%. This process is repeated 20 times. The parameter α in SSFCM is set to 1. λ_1, λ_2, and p in CS3FCM are respectively set to 1, 10 and 5. σ is set to the average distance between the samples.

In the experiments, we employ the clustering accuracy (CA) to measure the behavior of different methods. Formally, y_k and $\widetilde{y}_k$ respectively denote the ground-truth and obtained label of x_k. The CA for the unsupervised clustering methods is computed as:

$$\text{CA} = \frac{\sum_{k=1}^{n}\delta\left(y_k, \text{map}(\widetilde{y}_k)\right)}{n}$$

where $\delta(x, y)$ is the delta function that equals 1 if $x = y$ and 0, otherwise. $\text{map}(\widetilde{y}_k)$ is the permutation mapping function that maps the label $\widetilde{y}_k$ to the equivalent one using the Kuhn–Munkres algorithm (Lovasz and Plummer, 1986).

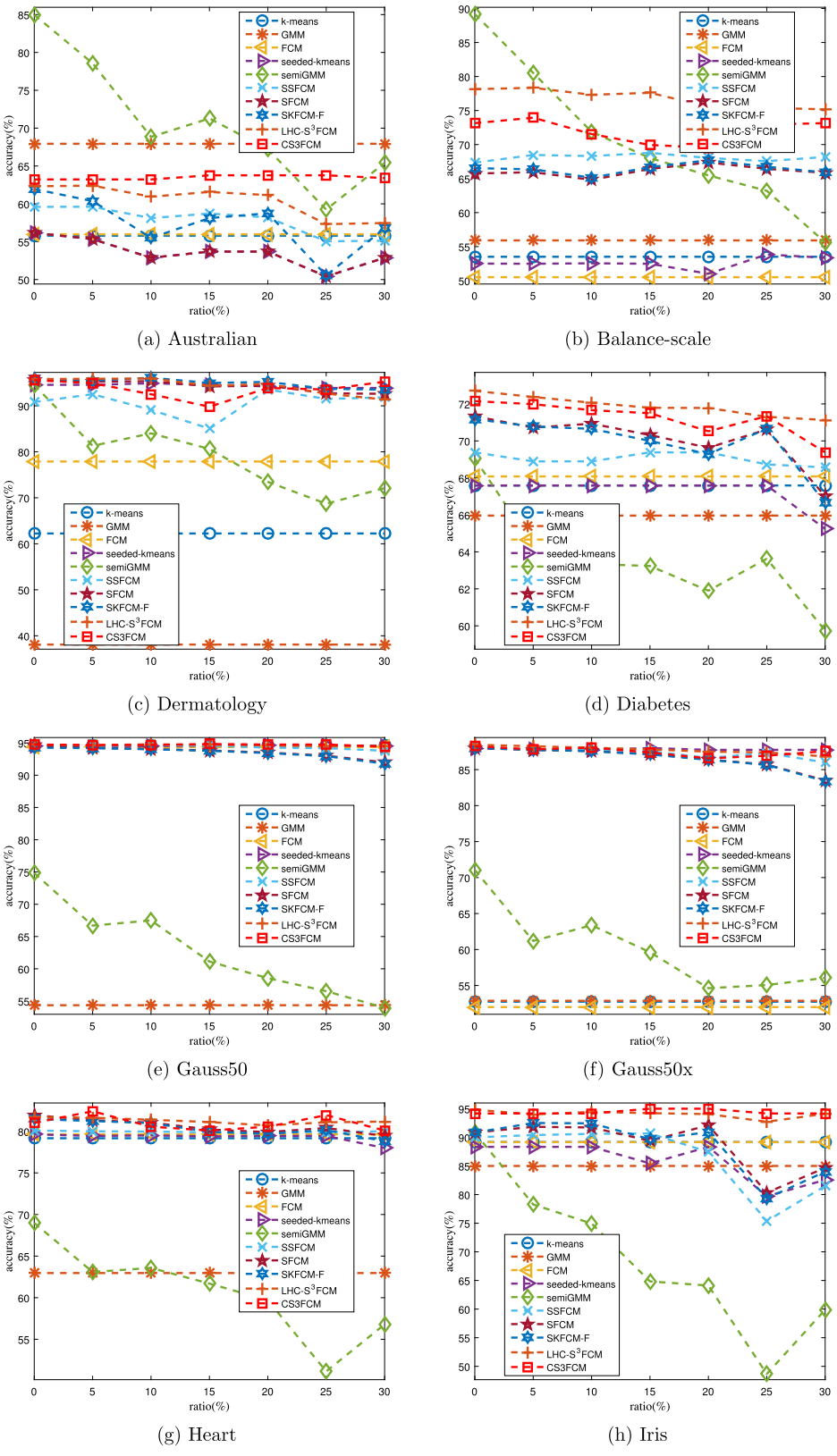

Fig. 2. Performance comparison of different methods over the first eight datasets.

The CA for the semi-supervised clustering methods is computed as:

$$\mathrm{CA} = \frac{\sum_{k=1}^{n} \delta\left(y_k, \widetilde{y}_k\right)}{n}$$

4.1. Result discussion

The accuracies of the different clustering methods as the different wrong ratios change are reported in Figs. 2–3. From these figures, we can have the following conclusions:

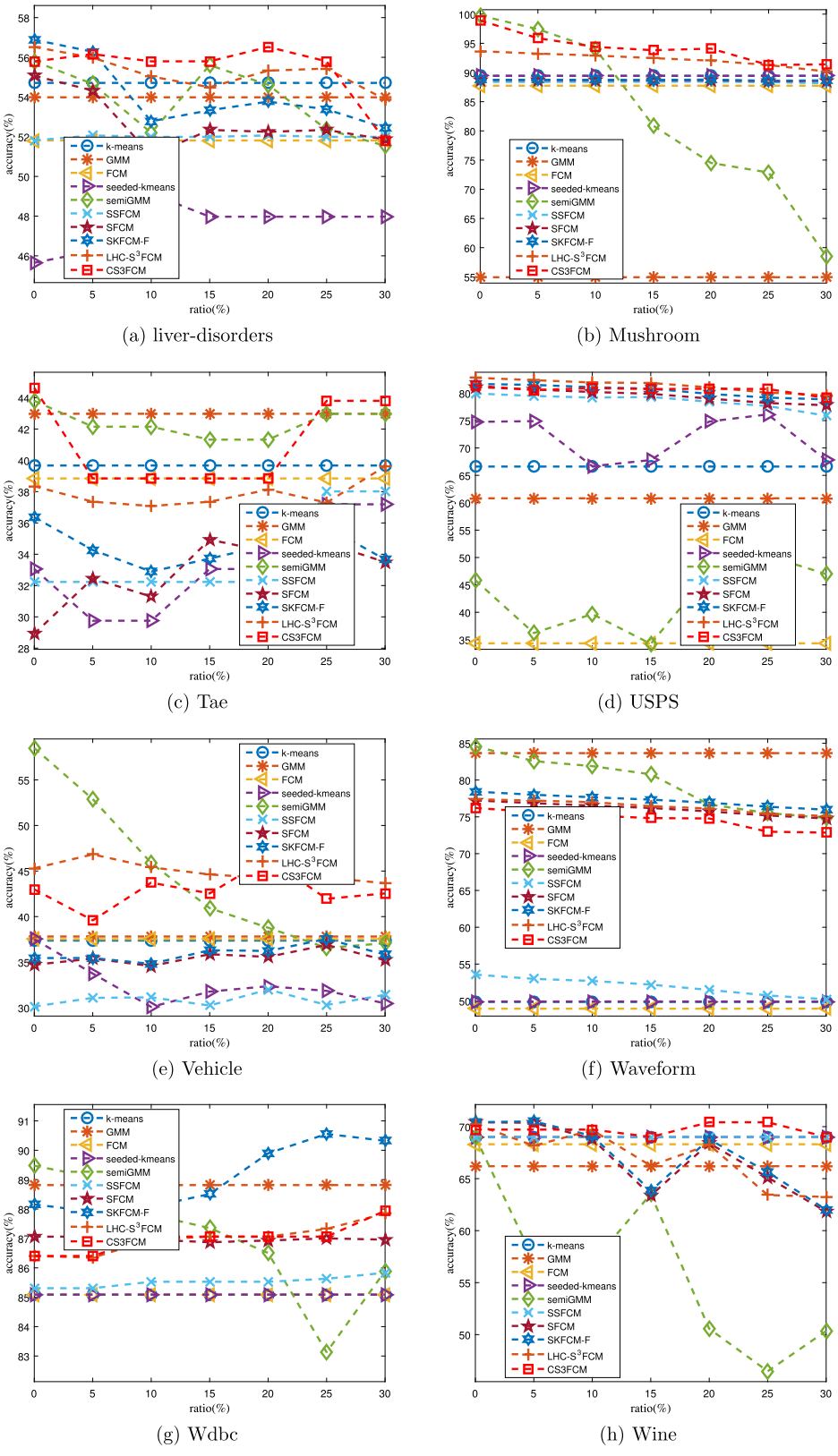

Fig. 3. Performance comparison of different methods over the latter eight datasets.

1. When the ratio is 0% (i.e., there are not wrong labels), SSFCM, SFCM and SKFCM-F can outperform FCM over the most datasets except Tae and Vehicle. It indicates that the prior knowledge may help improve the clustering performance.

2. When the ratio is 0%, CS3FCM performs better than unsupervised clustering (i.e., *k*-means, GMM and FCM) in most cases. It shows that our algorithm can be designed for semi-supervised clustering.

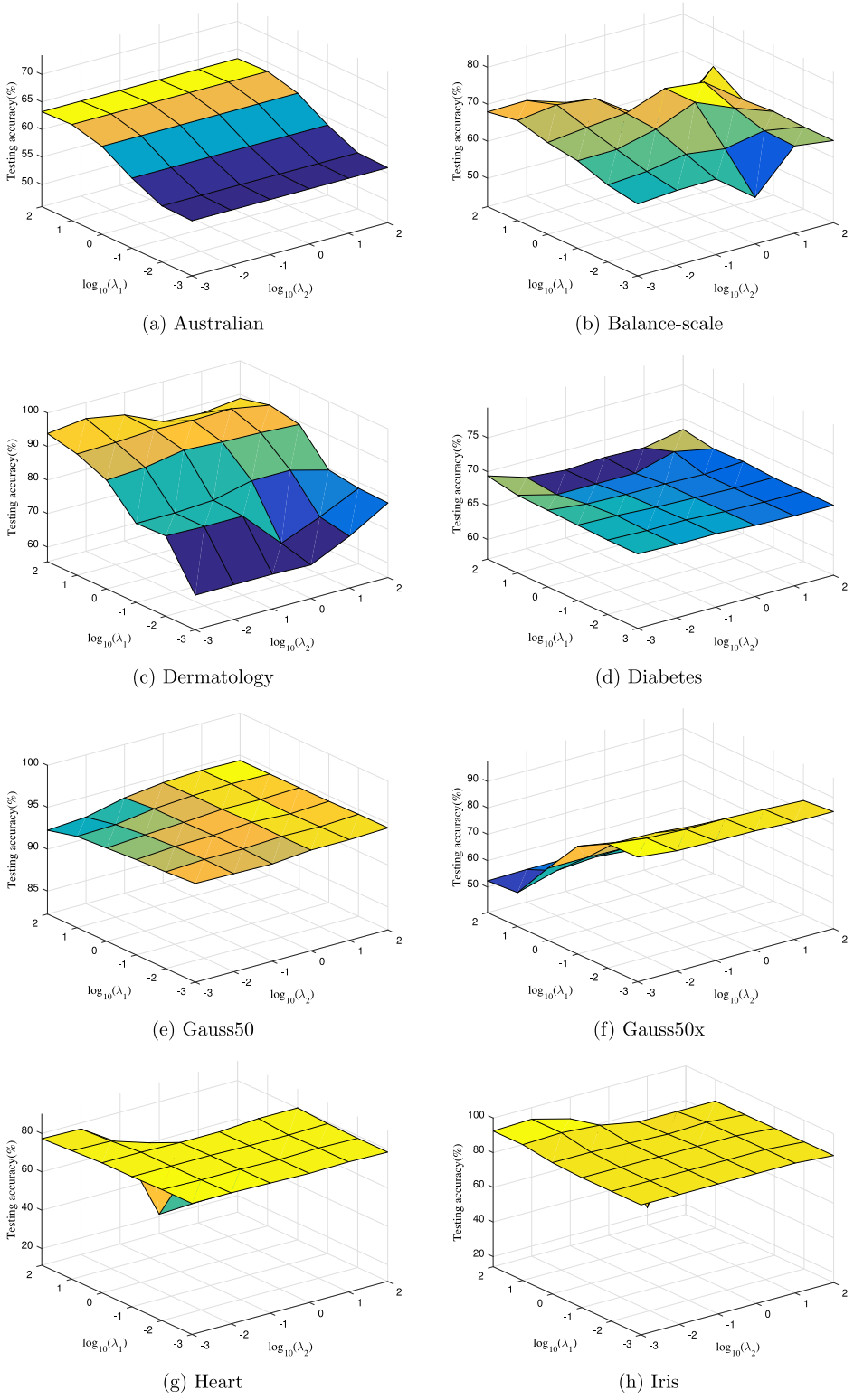

Fig. 4. Clustering performance with different regularization parameters over the first eight datasets.

3. The performance of SSFCM, SFCM, SKFCM-F, LHC-S^{3}FCM and our algorithm overall decreases as the wrong ratio increases in most datasets. It demonstrates that the wrong labels can hurt the performance of semi-supervised clustering. Meanwhile, one can see that CS3FCM achieves better results as the wrong ratio increases in some datasets, such as Heart, Tae, Vehicle and Wdbc. The reason may be that there is not equivalence between the clusters and classes. In this case the mislabeled samples may help to improve the clustering performance.
4. Semi-supervised clustering cannot always outperform unsupervised clustering. In some cases, SSFCM performs worse than FCM when the ratio reaches a certain value, such as more than 25% on Australian and 20% on IRIS. SFCM and SKFCM-F achieve worse performance than FCM when the ratio reaches a certain value, such as more than 20% on both Australian and IRIS,

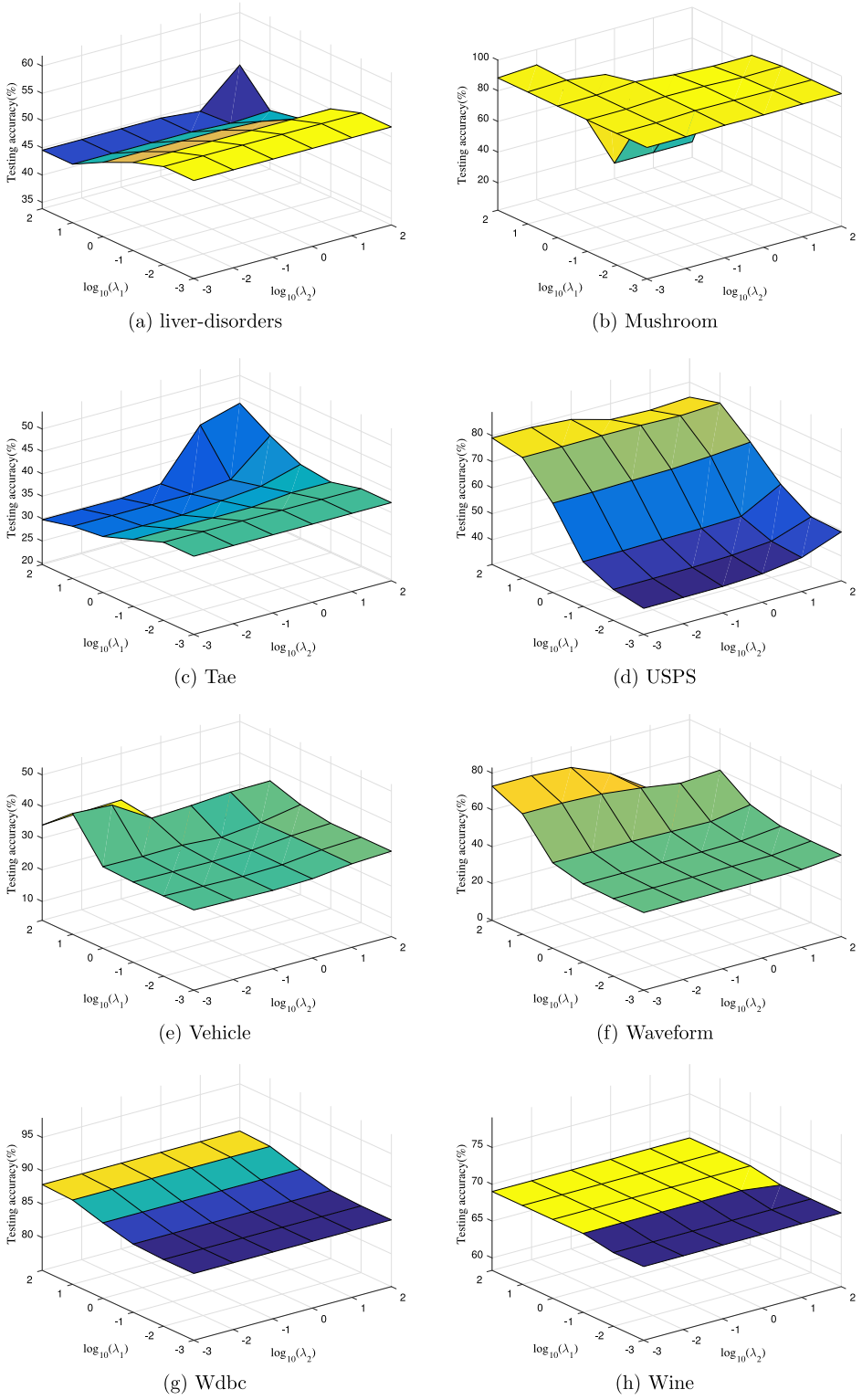

Fig. 5. Clustering performance with different regularization parameters over the latter eight datasets.

10% on Wine. More specially, FCM always outperforms SSFCM, SFCM and SKFCM-F under the different wrong ratios on Tae and Vehicle. These results further prove that the inappropriate prior knowledge is harmful for the clustering performance and it explains the necessary of designing safe semi-supervised clustering.

5. Compared to FCM and SSFCM, CS3FCM can obtain the best clustering performance in all cases. It indicates that the designed mechanism in our algorithm is effective and our algorithm can safely exploit the labeled samples.
6. In most cases, our algorithm can obtain comparable, if not better, than LHC-S^{3}FCM, especially on Australian, Mushroom, Tae and Wine. However, our algorithm performs worse than LHC-S^{3}FCM on Balance-scale, Vehicle and Waveform. This phenomenon may be related to the performance of FCM which is an important step for estimating the confidence weights. When

Table 1
Description of the experimental datasets.

Dataset	#Samples	#Features	#Classes
Australian	690	15	2
Balance-scale	625	4	3
Dermatology	336	33	6
Diabetes	768	8	2
Gauss50	1550	50	2
Gauss50x	2000	50	2
Heart	270	13	2
Iris	150	4	3
Liver-disorders	345	6	2
Mushroom	8124	112	2
Tae	151	5	3
USPS	2007	256	10
Vehicle	846	18	4
Waveform	5000	21	3
Wdbc	569	30	2
Wine	178	13	3

FCM achieves the desired results, our algorithm can obtain the appropriate weights and outperform LHC-S^{3}FCM. Otherwise, our algorithm may be inferior to LHC-S^{3}FCM.

7. It is pointed out that GMM can obtain the best performance compared to *k*-means and FCM, such as Australian, Tae, Waveform and Wdbc. Meanwhile, the performance of semiGMM is sensitive to the wrong ratio. Hence, it is meaningful to simultaneously perform the safe exploitation of prior knowledge and optimal selection of clustering methods.

4.2. Parameter analysis

In this section, we further analyze the impact of λ_1 and λ_2 on the performance of our algorithm. In order to explain the importance of the local structure as defined by the third term in Eq. (5), we conduct the experiment when the wrong ratio is set to 30%. The values of the two parameters are selected in $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$. Figs. 4–5 show the plots on the different datasets. From the plots, one can find that the best performance is generally obtained when the value of λ_2 is large. It explains that the proposed safe mechanism in our algorithm is effective and efficient.

5. Conclusion

This paper presents CS3FCM for safe semi-supervised clustering. CS3FCM assumes that the different samples should have the different safe confidences on the performance. And the confidences are estimated through FCM which is free from the wrong labels. The safe mechanism is implemented by balancing the tradeoff between the given labels and outputs of the local homogeneous unlabeled samples. The experimental results show that our algorithm can alleviate the harm of the wrongly labeled samples. In the future work, we mainly focus on the following aspects: (1) It is interesting to solve the optimization problem (5) when m is different from 2; (2) It is also meaningful to report the clustering performance with the other forms of the prior knowledge, such as pair-wise constraints; (3) How to simultaneously safely exploit the prior knowledge and select the optimal clustering method is another interesting topic. (4) Since SKFCM-F with Gaussian kernel can yield better performance than SFCM in most cases, it is worth designing the kernel version of our algorithm.

Acknowledgments

The work was supported by Zhejiang Provincial Natural Science Foundation of China under grant No. LY19F020040, and National Natural Science Foundation of China under grant No. 61601162, 61771178 and 61671197, and Zhejiang Provincial Natural Science Foundation of China under grant No. LY17F030021 and LY18F030009.

Declarations of interest

None.

References

de Amorim, R.C., Mirkin, B., 2012. Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering. Pattern Recognit. 45, 1061–1075.
Basu, S., Banerjee, A., Mooney, R.J., 2002. Semi-supervised clustering by seeding. In: Proceedings of the 19th International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 27–34.
Basu, S., Bilenko, M., Mooney, R.J., 2004. A probabilistic framework for semi-supervised clustering. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, USA, pp. 59–68.
Bensaid, A.M., Hall, L.O., Bezdek, J.C., Clarke, L.P., 1996. Partially supervised clustering for image segmentation. Pattern Recognit. 29, 859–871.
Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York, USA.
Bilenko, M., Basu, S., Mooney, R.J., 2004. Integrating constraints and metric learning in semi-supervised clustering. In: Proceedings of the Twenty-first International Conference on Machine Learning. ACM, New York, NY, USA, pp. 11–18.
Bouman, C.A., 1997. Cluster: An unsupervised algorithm for modeling Gaussian mixtures. Available from http://www.ece.purdue.edu/~bouman.
Chen, W., Feng, G., 2012. Spectral clustering: A semi-supervised approach. Neurocomputing 77, 229–242.
Chen, X., Liu, X., Jia, Y., 2011. Discriminative structure selection method of Gaussian Mixture Models with its application to handwritten digit recognition. Neurocomputing 74, 954–961.
Cohen, I., Cozman, F., Sebe, N., Cirelo, M., Huang, T., 2004. Semisupervised learning of classifiers: theory, algorithms, and their application to human–computer interaction. IEEE Trans. Pattern Anal. Mach. Intell. 26, 1553–1566.
Ding, S., Jia, H., Zhang, L., Jin, F., 2014. Research of semi-supervised spectral clustering algorithm based on pairwise constraints. Neural Comput. Appl. 24, 211–219.
Dong, A., Chung, F.-l., Wang, S., 2016. Semi-supervised classification method through oversampling and common hidden space. Inform. Sci. 349, 216–228.
Frank, A., Asuncion, A., 2010. UCI machine learning repository.
Gan, H., Fan, Y., Luo, Z., Zhang, Q., 2018. Local homogeneous consistent safe semi-supervised clustering. Expert Syst. Appl. 97, 384–393.
Gan, H., Luo, Z., Sun, Y., Xi, X., Sang, N., Huang, R., 2016. Towards designing risk-based safe laplacian regularized least squares. Expert Syst. Appl. 45, 1–7.
Gan, H., Sang, N., Chen, X., 2013a. Semi-supervised kernel minimum squared error based on manifold structure. In: Proceedings of the 10th International Symposium on Neural Networks, Vol. 7951. Springer-Verlag, Berlin, Heidelberg, pp. 265–272.
Gan, H., Sang, N., Huang, R., 2015. Manifold regularized semi-supervised gaussian mixture model. J. Opt. Soc. Amer. A 32, 566–575.
Gan, H., Sang, N., Huang, R., Tong, X., Dan, Z., 2013b. Using clustering analysis to improve semi-supervised classification. Neurocomputing 101, 290–298.
Givoni, I.E., Frey, B.J., 2009. Semi-supervised affinity propagation with instance-level constraints. J. Mach. Learn. Res.- Proc. Track 5, 161–168.
Hartigan, J.A., Wong, M.A., 1979. A K-means clustering algorithm. Appl. Stat. 28, 100–108.
Jain, A.K., 2010. Data clustering: 50 years beyond k-means. Pattern Recognit. Lett. 31, 651–666.
Jia, Y., Kwong, S., Hou, J., 2018. Semi-supervised spectral clustering with structured sparsity regularization. IEEE Signal Process. Lett. PP, 1–1.
Kalintha, W., Ono, S., Numao, M., Fukui, K.I., 2017. Kernelized evolutionary distance metric learning for semi-supervised clustering. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA. pp. 4945–4946.
Li, Y.-F., Kwok, J.T., Zhou, Z.-H., 2016a. Towards safe semi-supervised learning for multivariate performance measures. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. In: AAAI'16, AAAI Press, pp. 1816–1822.
Li, Y.-F., Wang, S.-B., Zhou, Z.-H., 2016b. Graph quality judgement: A large margin expedition. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. In: IJCAI'16, AAAI Press, pp. 1725–1731.
Li, Y., Zha, H., Zhou, Z., 2017. Learning safe prediction for semi-supervised regression. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA. pp. 2217–2223.
Li, Y.-F., Zhou, Z.-H., 2011a. Improving semi-supervised support vector machines through unlabeled instances selection. In: Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence. AAAI Press, pp. 500–505.
Li, Y.-F., Zhou, Z.-H., 2011b. Towards making unlabeled data never hurt. In: Proceedings of the 28th International Conference on Machine Learning. Omnipress, pp. 1081–1088.
Lovasz, L., Plummer, M., 1986. Matching Theory. North Holland, Budapest.
Mai, S.D., Ngo, L.T., 2015. Semi-supervised fuzzy c-means clustering for change detection from multispectral satellite image. In: 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). pp. 1–8.
Mai, S.D., Ngo, L.T., 2018. Multiple kernel approach to semi-supervised fuzzy clustering algorithm for land-cover classification. Eng. Appl. Artif. Intell. 68, 205–213.

Martinez-Uso, A., Pla, F., Sotoca, J.M., 2010. A semi-supervised gaussian mixture model for image segmentation. In: International Conference on Pattern Recognition, Vol. 0. IEEE Computer Society, Los Alamitos, CA, USA, pp. 2941–2944.

Pedrycz, W., Waletzky, J., 1997. Fuzzy clustering with partial supervision. IEEE Trans. Syst. Man Cybern. B 27, 787–795.

Singh, A., Nowak, R., Zhu, X., 2009. Unlabeled data: Now it helps, now it doesn't. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (Eds.), Advances in Neural Information Processing Systems 21. Curran Associates, Inc., pp. 1513–1520.

Wang, Y., Chen, S., 2013. Safety-aware semi-supervised classification. IEEE Trans. Neural Netw. Learn. Syst. 24, 1763–1772.

Wang, H., Wang, S.-B., Li, Y.-F., 2016. Instance selection method for improving graph-based semi-supervised learning. In: Booth, R., Zhang, M.-L. (Eds.), PRICAI 2016: Trends in Artificial Intelligence: 14th Pacific Rim International Conference on Artificial Intelligence, Phuket, Thailand, August 22-26, 2016, Proceedings. Springer International Publishing, Cham, pp. 565–573.

Wei, S., Li, Z., Zhang, C., 2017. Combined constraint-based with metric-based in semi-supervised clustering ensemble. Int. J. Mach. Learn. Cybern. 1–16.

Yan, Y., Chen, L., Nguyen, D.T., 2012. Semi-supervised clustering with multi-viewpoint based similarity measure. In: The 2012 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–8.

Yang, T., Priebe, C.E., 2011. The effect of model misspecification on semi-supervised classification. IEEE Trans. Pattern Anal. Mach. Intell. 33, 2093–2103.

Yin, X., Chen, S., Hu, E., Zhang, D., 2010. Semi-supervised clustering with metric learning: An adaptive kernel method. Pattern Recognit. 43, 1320–1333.

Yin, X., Shu, T., Huang, Q., 2012. Semi-supervised fuzzy clustering with metric learning and entropy regularization. Knowl.-Based Syst. 35, 304–311.

Zhang, H., Lu, J., 2009. Semi-supervised fuzzy clustering: A kernel-based approach. Knowl.-Based Syst. 22, 477–481.

Zhang, Y., Wen, J., Wang, X., Jiang, Z., 2015. Semi-supervised hybrid clustering by integrating gaussian mixture model and distance metric learning. J. Intell. Inf. Syst. 45, 113–130.