IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 3, 2019, accepted August 27, 2019, date of publication September 5, 2019, date of current version September 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2939581

Stratified Feature Sampling for Semi-Supervised Ensemble Clustering

JIALIN TIAN[1], YAZHOU REN[1], AND XIANG CHENG[2]
[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, USA

Corresponding author: Yazhou Ren (yazhou.ren@uestc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61806043, Grant 61832001, and Grant 61872062, and in part by the Project funded by China Postdoctoral Science Foundation under Grant 2016M602674.

ABSTRACT Ensemble Clustering (EC), which seeks to generate a consensus clustering by integrating multiple base clusterings, has attracted increasing attentions. However, traditional EC methods typically have three main limitations: (1) High dimensional data present a huge challenge to ensemble clustering methods. (2) Most EC algorithms can not use prior information, e.g., pairwise constraints, to enhance the clustering performance. (3) Even in existing semi-supervised ensemble clustering methods, prior information is not sufficiently used, e.g., only used in generating base clusterings. To alleviate these problems, we propose Stratified Feature Sampling for Semi-Supervised Ensemble Clustering (SFS^{3}EC). Firstly, we develop a novel stratified feature sampling method, which can cope with high dimensional data, guarantee the diversity of base clusterings, and reduce the risk that some features are not selected at the same time. Secondly, semi-supervised clustering, i.e., constraint propagation, is applied to obtain base clusterings. Finally, we propose to utilize prior information in both the base clustering generating process and the consensus process, which guarantees that prior information is sufficiently used. We conduct a series of experiments on ten real-world data sets to demonstrate the effectiveness of the proposed model.

INDEX TERMS Constraint propagation, ensemble clustering, high dimensional data, semi-supervised learning, stratified feature sampling.

I. INTRODUCTION

Clustering is an essential data analysis and visualization tool that has been used in a wide range of applications, including document analysis, image segmentation, and image retrieval [1], etc. Numerous clustering algorithms have been proposed in the past decades, including k-means [2], hierarchical clustering [3], DBSCAN [4], non-negative matrix factorization based clustering [5], etc. Different clustering methods provide different clustering results and it is hard to decide which result should be used. Besides, for high dimensional data, most traditional clustering methods fail in obtaining good performance due to sparsity, noise, and correlation of features.

To address the above issues, ensemble clustering [6] is proposed to improve the performance by making use of information from multiple base clusterings. More specifically, an ensemble clustering method is usually superior to a single clustering method in terms of robustness, consistency, and stability. Ensemble clustering can be typically divided into two steps, i.e., generating base clusterings and integrating these clusterings into a consensus one.

For high dimensional data, two randomization methods called random feature sampling [7], [8] and random projection [9]–[11] are commonly used to generate different feature subsets. However, these feature subsets do not well represent the characteristics of the original data because randomization methods can not explore the correlations among original features. To address this, stratified feature sampling (SFS) [12] is proposed. Firstly, SFS uses k-means to partition features into several feature groups. Then, it randomly selects features from each feature group with a same proportion to form a feature subset. SFS has the risk that some features will not be selected to generate the base clusterings. In this work, we develop a novel stratified feature sampling method to reduce such risk.

From another perspective, pairwise constraints can be used to effectively improve the performance of traditional clustering algorithms. There are usually two types of pairwise constraints, i.e., must-link constraints and cannot-link constraints, which respectively indicate whether two data points should be assigned to the same cluster or not. There are a

The associate editor coordinating the review of this manuscript and approving it for publication was Gianluigi Ciocca.

number of semi-supervised clustering algorithms that have been proposed [13]–[15]. Many of existing methods only use pairwise constraints to generate base clusterings. By contrast, our proposed model makes use of prior information in both the base clustering generating process and the consensus process.

In this paper, we propose an efficient semi-supervised ensemble clustering model, namely Stratified Feature Sampling for Semi-Supervised Ensemble Clustering (SFS^{3}EC). Firstly, a novel stratified feature sampling method is proposed to generate feature subsets. Secondly, these feature subsets are used to generate a set of base clusterings through the constraint propagation method [13]. Finally, these base clusterings are integrated into the consensus clustering. It is worthy to notice that pairwise constraints are sufficiently utilized in both the process of base clusterings generation and the process of ensemble clustering. Extensive experiments are conducted on real-world data sets to show the effectiveness of SFS^{3}EC.

The contributions of this paper are summarized as below:

1. A novel stratified feature sampling method is proposed to reduce the risk that a part of features can not participate in the clustering process.
2. Pairwise constraints are adequately used in both the process of generating base clusterings and the process of ensemble clustering to enhance the clustering performance.
3. Extensive experiments demonstrate the effectiveness and efficiency of the proposed model. In addition, parameter sensitivity analysis of the proposed model is also provided.

The rest of paper is arranged as follows: Section 2 introduces the related work. Section 3 illustrates the proposed SFS^{3}EC in detail. Section 4 describes the experimental settings and Section 5 shows experimental results and the corresponding analysis. Section 6 gives the conclusion and further work.

II. RELATED WORK

Clustering is a technique that divides multi-dimensional data into closely related clusters. In the past few decades, a lot of clustering algorithms have been proposed, such as k-means [2], DBSCAN [4], hierarchical clustering [3], mean shift clustering [16]–[18], unsupervised deep embedding clustering [19], non-negative matrix factorization based clustering [20], multi-view spectral clustering [21], etc.

Ensemble clustering integrates multiple clustering results into a single clustering result, with recognized advantages in generating robust partitions and handling noises. Since the framework of ensemble clustering was formalized by [6], many different ensemble clustering algorithms have been proposed, such as hierarchical ensemble clustering [22], divisive clustering ensemble with automatic cluster number [23], weighted-object ensemble clustering [24], [25], spectral ensemble clustering based on co-association matrix [26], stratified feature sampling ensemble clustering [12], locally weighted ensemble clustering [27], spectral ensemble clustering via weighted k-means [28], multiple kernel fuzzy clustering [29], and so on.

Semi-supervised clustering utilizes a small mount of prior knowledge such as pairwise constraints to enhance the clustering performance. There are many semi-supervised clustering algorithms have been proposed, e.g., constraints k-means [30], PCKmeans [15], C-DBSCAN [31], semi-supervised maximum margin clustering [32], exhaustive and efficient constraint propagation [13], semi-supervised deep embedded clustering [14], semi-supervised denpeak clustering with pairwise constraints [33], semi-supervised deep fuzzy c-mean clustering [34], etc.

However, the performance of semi-supervised clustering methods is still not robust and stable, since it is sensitive to the value of parameters and the order of pairwise constraints. To address this, semi-supervised ensemble clustering, which can further improve the robustness, stability and accuracy of clustering results, has been emerged recently. Representative methods includes incremental semi-supervised clustering ensemble [35], private aggregation of teacher ensembles [36], temporal ensemble for semi-supervised learning [37], adaptive ensembling of semi-supervised clustering solutions [38], semi-supervised ensemble clustering based on selected constraint projection [39], etc.

III. PROPOSED APPROACH

In this section, we will elucidate the framework of Stratified Feature Sampling for Semi-Supervised Ensemble Clustering (SFS^{3}EC) in detail. Firstly, a novel stratified sampling method is proposed to generate several subsets of features. The subsets are denoted as $S_1, S_2, \ldots, S_r$, where r is the number of feature subsets and is equal to the number of base clusterings as well. Secondly, based on spectral clustering, the constraint propagation is used to propagate pairwise constraints to the whole data set and generates a number of base clusterings, which are denoted as $C_1, C_2, \ldots, C_r$. Thirdly, a similarity matrix is constructed by base clusterings and then is adjusted by the constraint propagation again. Finally, the similarity matrix is partitioned by METIS algorithm [40] to get the final clustering result. Fig. 1 shows the process of SFS^{3}EC.

A. STRATIFIED FEATURE SAMPLING

Given a data set $X = \{x_1, x_2, \ldots, x_n\}$, where x_i denotes an m-dimensional data point. We propose a novel stratified feature sampling technique which firstly divides the m features into c clusters by k-means. Then, features are selected from each feature cluster individually to construct a feature subset. The sampling ratio is set to the same value for all feature groups. Suppose the sampling ratio p is equal to 0.2, then 20% of features in each feature cluster are selected. Fig. 2 shows the overview of stratified feature sampling.

If all features always have the equal possibility to be chosen, it is of high risk that some features are sampled multiple times while others will not be selected in all the feature

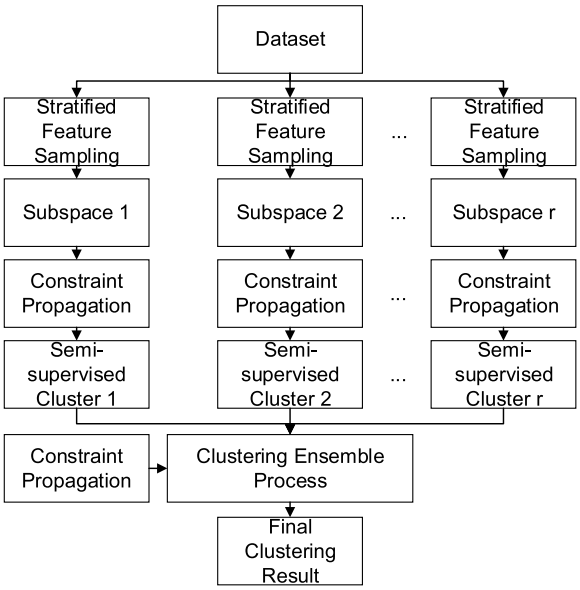


FIGURE 1. Overview of stratified feature sampling for semi-supervised ensemble clustering.

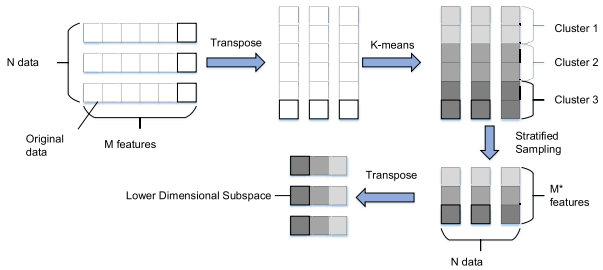


FIGURE 2. Overview of stratified feature sampling.

subsets. To alleviate this problem, we design a novel stratified feature sampling method which changes the sampling probability of each feature according to the sampling history. To be specific, when constructing the first feature subset, every feature's probability to be sampled is initialized to be equal. If some features are selected in the current feature subset, their sampling probabilities will be halved in the next turn.

Fig. 3 gives an example of how to change the sampling probabilities in a specific feature cluster consisted of 10 features. The size of each feature subset is 5. The numbers in boxes is the probabilities for features to be sampled. The blackened boxes indicate that corresponding features are selected. At first, the probabilities are initialized equally to be 10%. Five features are randomly sampled to form the first feature subset, and the corresponding probabilities are reduced to 5%. Then, all features' probabilities are normalized to add up to 1. This procedure is repeated in the second and subsequent stratified sampling processes. After repeating it r times, all the feature subsets are generated.

Given an m-dimensional data set, let the sampling ratio be p and the number of feature subsets be r. If all features are selected with the same probability, it is of probability $p^* = (1 - \frac{1}{m})^{rmp}$ that a feature is not selected in all the r sampling rounds, and the expectation of the number of features that are not selected is $m \cdot p^*$. For example, if $p = 0.3, m = 1000$, and

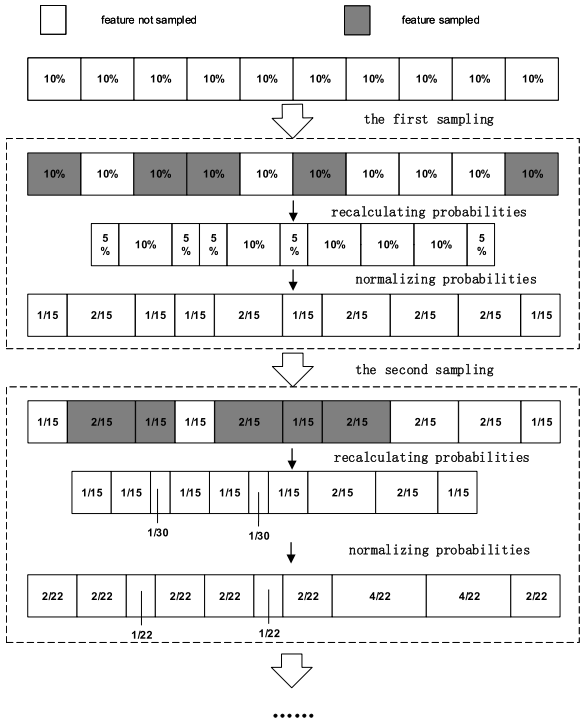


FIGURE 3. An example of stratified feature sampling.

$r = 10$, it is expected that 49.71 features will not have any impact on the clustering results.

By contrast, according to our feature sampling strategy, the features that are not selected in many successive rounds will be the most likely being selected of all the features in the next round. It is expected that the number of unselected features in all r rounds will be reduced.

To verify this, we have done experiments to simulate these two sampling methods. Here, we set $p = 0.3$, $m = 1000$, and $r = 10$. The two sampling methods are applied for 100 independent runs. The average number of unselected features in all the 10 sampling rounds is recorded. The average number of unselected features of the random sampling method is 48.85, while that of our method is only 4.17, which is greatly reduced.

B. GENERATING BASE CLUSTERINGS

1) CONSTRAINT MATRIX

A form of commonly used semi-supervised information is called pairwise constraints. Must-link constraints and cannot-link constraints are two types of pairwise constraints, which indicate whether two data points should be assigned to the same cluster or not respectively. We define a set of must-link constraints as $M = \{(x_i, x_j) : y_i = y_j, 1 \le i, j \le n\}$ and a set of cannot-link constraints as $C = \{(x_i, x_j) : y_i \ne y_j, 1 \le i, j \le n\}$, where y_i is the ground-truth label of x_i. Then, the constraint matrix $O = \{o_{ij}\}_{n \times n}$ is defined as:

$$O = \begin{cases} +1, & (x_i, x_j) \in M \\ -1, & (x_i, x_j) \in C \\ 0, & otherwise \end{cases} \quad (1)$$

2) SIMILARITY MATRIX

Then, we define a similarity matrix as $E = \{e_{ij}\}_{n*n}$ for every feature subset, where e_{ij} represents the similarity between x_i and x_j. It is formulated as:

$$e_{ij} = \begin{cases} l_{ij}, & if\ x_i \in N_k(x_j)\ or\ x_j \in N_k(x_i) \\ 0, & otherwise \end{cases} \tag{2}$$

$$l_{ij} = exp(-\frac{\|x_i - x_j\|^2}{\overline{d}^2}) \tag{3}$$

where $N_k(x_i)$ denotes the k-nearest neighbours of x_i and $\overline{d}$ is the average Euclidean distance from every point to its k-nearest neighbor.

3) CONSTRAINT PROPAGATION

Constraint propagation [13] utilizes pairwise constraints to adjust the similarity matrix E for each base clustering. Then, the adjusted similarity matrix is used in spectral clustering. First, L is defined as a normalized graph Laplacian of E. Second, a propagation matrix is defined as $F = \{f_{ij}\}_{n\times n}$, where $|f_{ij}| \leq 1$. F is initialized as the constraint matrix O.

Constraint propagation can be divided into vertical propagation and horizontal propagation. According to [13], constraint propagation can be performed in two directions at the same time as following:

$$\min_F \frac{1}{2}\|F - O\|_F^2 + \frac{\mu}{2} tr(F^T L F + F L F^T) \tag{4}$$

where $\mu > 0$ is a regularization parameter, tr is the trace of a matrix.

The closed-form solution $F^* = \{f_{ij}^*\}_{n\times n}$ is formulated as follows [13]:

$$F^* = (1-\beta)^2(1-\beta\overline{L})^{-1} O (I - \beta\overline{L})^{-1} \tag{5}$$

where $\beta = \mu/(\mu+1)$ and $\overline{L} = I - L$.

Then, F^* is normalized by $\widetilde{f}_{ij}^* = \frac{f_{ij}^*}{max_{i',j'}|f_{i'j'}^*|}$. $\widetilde{F}^*$ will be further used to adjust the similarity matrix E as follows:

$$e_{ij} = \begin{cases} 1 - (1 - \widetilde{f}_{ij}^*)(1 - e_{ij}), & \widetilde{f}_{ij}^* \geq 0 \\ (1 + \widetilde{f}_{ij}^*) e_{ij}, & \widetilde{f}_{ij}^* < 0 \end{cases} \tag{6}$$

Finally, spectral clustering (which is normalized according to Shi and Malik [41]) makes use of the adjusted similarity matrix E to generate a base clustering. In this way, all the r base clusterings $C_1, C_2, ..., C_r$ can be generated.

C. CONSENSUS CLUSTERING WITH CONSTRAINT PROPAGATION

After getting r base clusterings, we need to incorporate them into one final clustering result. Firstly, a hypergraph H is built.

When a hypergraph is constructed, a similarity matrix R could be constructed as:

$$R = \frac{1}{r} H H^T \tag{7}$$

where r is the number of base clusterings.

Matrix R contains all the information uncovered by different single clustering results. Traditionally, R is directly used to produce the final ensemble clustering. By contrast, in our method, R is refined with prior knowledge, i.e., pairwise constraints. Concretely, we use constraint propagation to obtain a refined R with pairwise constraints. Then, a well-known graph partition algorithm called METIS is applied to partition the adjusted similarity matrix R to generate our final clustering result.

IV. EXPERIMENTAL SETUP

A. DATA SETS

We evaluate the performance of the comparing methods on 10 real-world data sets. Table 1 gives an overview of these data sets, where AustralianCredit, Biodeg (QSAR biodegradation), CNAE-9, Iris, and Protein are from UCI machine learning repository.[1] Brain and Colon are gene expression data sets.[2] ORL-32 × 32 and Yale-32 × 32 are popular face databases.[3] TwoLeadECG is a time series data set.[4] Every feature is normalized to have zero mean value and unit variance. The ground-truth labels of each data set are used to generate pairwise constraints and to evaluate the performance of clustering algorithms.

TABLE 1. Data sets used in the experiment (n is the number of data points and m is the number of features).

Data sets	n	m	classes
AustralianCredit	690	14	2
Biodeg	1055	41	2
Brain	42	5597	5
CNAE-9	1080	856	9
Colon	62	2000	2
Iris	150	4	3
ORL-32x32	400	1024	40
Protein	116	20	6
TwoLeadECG	1162	82	2
Yale-32x32	165	1024	15

B. COMPARING METHODS

The following clustering methods are tested in the experiments:

1. **k-means**: k-means clustering [2].
2. **SFSEC**: Stratified feature sampling ensemble clustering [12], which uses stratified feature sampling (SFS) to generate base clusterings, and then adopts three consensus functions called CSPA, MCLA and HBGF [6] to generate the final clustering result.
3. **E²CP**: Exhaustive and efficient constraint propagation [13], which is a single semi-supervised clustering algorithm with constraint propagation.
4. **ISSCE**: Incremental semi-supervised clustering ensemble [35], a semi-supervised clustering ensemble algorithm, which designs an incremental ensemble

[1] https://archive.ics.uci.edu/ml/index.php.
[2] https://stat.ethz.ch/~dettling/bagboost.html.
[3] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html.
[4] https://www.cs.ucr.edu/~eamonn/time_series_data/.

TABLE 2. Comparison against multiple methods with respected to NMI.

Data sets	k-means	SFSEC			E^{2}CP	ISSCE	SFS^{3}EC*	SFS^{3}EC
		CSPA	HGPA	MCLA				
AustralianCredit	0.3362	0.2935	0.0014	0.2424	0.1681	0.4251	**0.4590**	**0.4553**
Biodeg	0.1082	0.1107	0.0002	0.1259	0.1721	0.2273	0.2536	**0.3087**
Brain	0.4686	**0.5464**	0.4627	0.5170	0.4954	0.5167	**0.5329**	**0.5520**
CNAE-9	0.1184	0.1756	0.0692	0.0169	0.0861	0.0468	0.0860	**0.2829**
Colon	0.0400	0.0096	0.0326	0.0094	0.1030	0.0737	**0.1507**	**0.1677**
Iris	0.6486	0.8497	0.3235	0.8248	0.6624	**0.8548**	**0.8695**	**0.8674**
ORL-32x32	0.7410	0.7748	0.7778	0.7847	0.8220	**0.8397**	0.8286	**0.8467**
Protein	0.3799	0.3654	0.3200	0.3630	0.2795	0.5030	0.5034	**0.5400**
TwoLeadECG	0.0004	0.0009	0.0011	0.0007	0.7888	**0.8895**	**0.8930**	**0.8914**
Yale-32x32	0.5154	0.5631	0.5539	0.5462	**0.5935**	**0.6081**	**0.5964**	**0.6069**

TABLE 3. Comparison against multiple methods with respected to ARI.

Data sets	k-means	SFSEC			E^{2}CP	ISSCE	SFS^{3}EC*	SFS^{3}EC
		CSPA	HGPA	MCLA				
AustralianCredit	0.4139	0.3616	0.0004	0.3096	0.2031	0.5143	**0.5502**	**0.5466**
Biodeg	0.0361	0.1253	-0.0006	-0.0171	0.2350	0.2417	0.2702	**0.3252**
Brain	0.3075	**0.4048**	0.3089	0.3773	0.3105	**0.4064**	**0.4034**	**0.4171**
CNAE-9	0.0092	0.1128	0.0340	0.0017	0.0101	0.0163	0.0604	**0.2130**
Colon	0.0186	-0.0043	0.0249	-0.052	0.1130	0.0753	**0.1546**	**0.1714**
Iris	0.5942	0.8631	0.2902	0.8241	0.6714	**0.8759**	**0.8917**	**0.8898**
ORL-32x32	0.3600	0.4757	0.4666	0.4773	0.4609	**0.6131**	0.5835	**0.6320**
Protein	0.2272	0.2365	0.1951	0.2262	0.1489	0.3966	0.3892	**0.4294**
TwoLeadECG	-0.0003	0.0004	0.0007	0.0001	0.8435	**0.9411**	**0.9442**	**0.9427**
Yale-32x32	0.2378	0.3141	0.3097	0.2917	0.3232	**0.3828**	**0.3608**	**0.3715**

member selection process to remove redundant ensemble members.

5. **SFS^{3}EC***: It is almost identical to the proposed SFS^{3}EC except that pairwise constraints are not used in the consensus step.
6. **SFS^{3}EC**: The proposed stratified feature sampling for semi-supervised ensemble clustering method.

C. PARAMETERS SETTING

All of the experiments are implemented on a 64-bit Microsoft Windows machine with 8 GB memory and Intel Core i5-8250U CPU of 1.60 GHz processing speed. Except for SFSEC and k-means that do not use pairwise constraints, all the other algorithms use the same pairwise constraints in an independent run. The ratio of feature sampling (p) is set to [0.1, 0.2, 0.3, 0.4, 0.5]. According to [35], the k of the k-nearest neighbor is set to 10. The number of feature clusters c is set to $\sqrt{m}$, where m is the number of total features. The number of pairwise constraints n_c ranges from $0.2n$ to $2n$ with each increment set to $0.2n$, where n is the number of data points. The number of base clusterings r is set to [10, 20, 30, 40, 50] and the cluster number of each base clustering as well as the final clustering is set to the number of ground-truth classes. Every algorithm is repeated 10 times and the average performance evaluation will be reported.

D. CLUSTERING EVALUATION METRICS

To evaluate the performance of clustering algorithms, normalized mutual information (NMI) [42] and adjusted rand index (ARI) [42] are adopted as evaluation metrics. The NMI and ARI values are in ranges [0, 1] and [−1, 1], respectively. The lager the two values are, the better the clustering results are. t-test is used to assess the statistical significance of the results at 5% significance level.

V. RESULTS AND ANALYSIS

A. RESULTS ON REAL DATA

We set the number of pairwise constraints to n, the number of base clusterings to 20, and the ratio of feature sampling to 0.3. Table 2 and Table 3 show the clustering results w.r.t. NMI and ARI, respectively. In each row of Table 2 and Table 3, the best and comparable results are highlighted in boldface. Table 4 evaluates the time complexity of our method, where the smaller running time is highlighted in boldface in each row.

TABLE 4. The comparison of average execution time (second) between SFS^{3}EC and ISSCE (the better value is highlighted in boldface).

Data sets	SFS^{3}EC	ISSCE
AustralianCredit	**14.96**	45.14
Biodeg	**40.22**	115.38
Brain	**2.37**	6441.33
CNAE-9	**53.16**	907.33
Colon	**1.12**	126.79
Iris	**0.74**	92.24
ORL-32x32	**7.49**	5577.42
Protein	**0.86**	67.48
TwoLeadECG	**52.55**	261.66
Yale-32x32	**2.06**	1211.96

Several interesting observations can be obtained from Tables 2 - 4. 1) No matter which consensus function is used

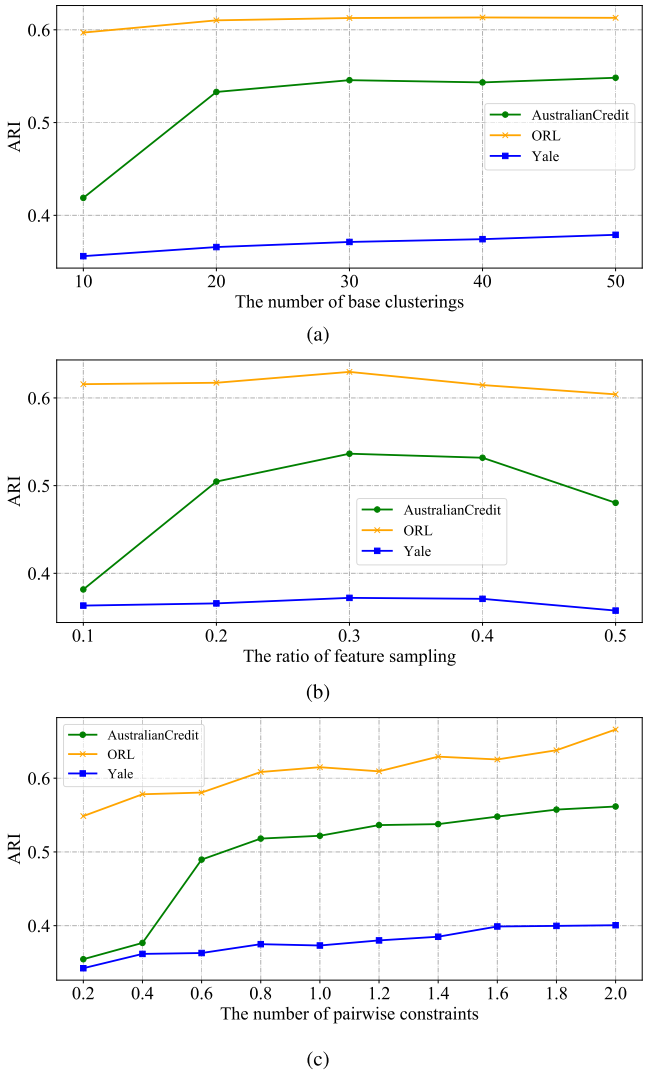


FIGURE 4. Sensitivity analysis of (a) the number of base clusterings, (b) the ratio of feature sampling, and (c) the number of pairwise constraints.

by SFSEC, its performance is relatively poor. This indicates that the semi-supervised information could significantly promote the clustering performance. 2) k-means and E^{2}CP generally lose to ISSCE and SFS^{3}EC, verifying the effectiveness of ensemble clustering. 3) Both our proposed SFS^{3}EC and ISSCE are semi-supervised ensemble clustering methods which utilize the prior information. However, ISSCE is much more time consuming than SFS^{3}EC, as shown in Table 4. In addition, according to Table 2 and 3, the performance of SFS^{3}EC is always better than or comparable with that of ISSCE. 4) SFS^{3}EC* is capable of generating clustering results with sufficient performance on most data sets. SFS^{3}EC can further improve the performance of SFS^{3}EC*, especially on Biodeg, CNAE-9, and Protein. It indicates that the usage of pairwise constraints in the consensus process could improve the performance of semi-supervised ensemble clustering. In summary, SFS^{3}EC generally generates the best clustering results in an efficient way.

B. SENSITIVITY ANALYSIS

In this section, we analyze the sensitivity of SFS^{3}EC w.r.t. the number of pairwise constraints (n_c), the number of base clusterings (r), and the ratio of feature sampling (p) on AustralianCredit, ORL-32 $\times$ 32, and Yale-32 $\times$ 32. The corresponding results are shown in Fig. 4.

The sensitivity analysis of r is tested with n_c and p set to n and 0.3. It is shown from Fig. 4(a) that the performance of SFS^{3}EC generally grows as r increases in the beginning. Then, the performance becomes stable when r is large enough, i.e., $r > 20$. Considering the trade-off between time complexity and performance, it is suggested that r should be set in range [20, 30].

Then, we test the sensitivity of the ratio of feature sampling p when $n_c = n$ and $r = 20$. The results are given in Fig. 4(b). SFS^{3}EC performs stably in a wide range of p on ORL-32 $\times$ 32 and Yale-32 $\times$ 32. On AustralianCredit, the performance grows firstly as p increases to 0.3 and declines as p continues to increase. The main reason is that a larger number of features for each base clustering could enhance its performance. However, as this number continues to increase, the diversity of base clusterings is reduced and noisy features might participate more in the clustering process, leading to negative impact on the ensemble clustering performance. As a consequence, the recommended value for p is 0.3.

Finally, we analyze the sensitivity of n_c with r and p set to 20 and 0.3, respectively. As shown in Fig. 4(c), the clustering performance of SFS^{3}EC generally becomes better as n_c increases on these three tested data sets, showing that more prior information is more beneficial to the performance of semi-supervised clustering.

VI. CONCLUSION AND FUTURE WORK

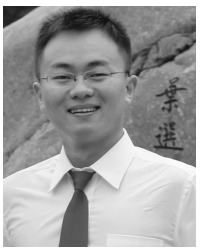
In this paper, we propose a model named stratified feature sampling for semi-supervised ensemble clustering (SFS^{3}EC), which develops a novel stratified feature sampling method, and incorporates pairwise constraints into both the base clusterings generating process and the consensus clustering process. The experiments demonstrate that our algorithm can stably generates satisfied clustering results in an efficient way. To exploit instance sampling or weighting strategy into semi-supervised ensemble clustering is an interesting future work.

REFERENCES

[1] Y. Ren, N. Wang, M. Li, and Z. Xu, "Deep density-based image clustering," Dec. 2018, *arXiv:1812.04287*. [Online]. Available: https://arxiv.org/abs/1812.04287
[2] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, 1967, pp. 281–297.
[3] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999.
[4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, 1996, pp. 226–231.
[5] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 12, 2001, pp. 556–562.

[6] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, Dec. 2002.
[7] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," *ACM SIGMOD Rec.*, vol. 27, no. 2, pp. 94–105, Jun. 1998,
[8] J. G. Dy and C. E. Brodley, "Feature subset selection and order identification for unsupervised learning," in *Proc. ICML*, 2000, pp. 247–254.
[9] K. Fukunaga, *Statistical Pattern Recognition*. New York, NY, USA: Academic, 1990.
[10] K. Chakrabarti, E. Keogh, S. Mehrotra, and M. Pazzani, "Locally adaptive dimensionality reduction for indexing large time series databases," *ACM Trans. Database Syst.*, vol. 27, pp. 188–228, Jun. 2002.
[11] X. Z. Fern and C. E. Brodley, "Random projection for high dimensional data clustering: A cluster ensemble approach," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2003, pp. 186–193.
[12] L. Jing, K. Tian, and J. Z. Huang, "Stratified feature sampling method for ensemble clustering of high dimensional data," *Pattern Recognit.*, vol. 48, no. 11, pp. 3688–3702, 2015.
[13] Z. Lu and Y. Peng, "Exhaustive and efficient constraint propagation: A graph-based learning approach and its applications," *Int. J. Comput. Vis.*, vol. 103, no. 3, pp. 306–325, 2013.
[14] Y. Ren, K. Hu, X. Dai, L. Pan, S. C. H. Hoi, and Z. Xu, "Semi-supervised deep embedded clustering," *Neurocomputing*, vol. 325, pp. 121–130, Jan. 2019.
[15] S. Basu, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering," in *Proc. SIAM Int. Conf. Data Mining*, 2004, pp. 333–344.
[16] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.
[17] Y. Ren, C. Domeniconi, G. Zhang, and G. Yu, "A weighted adaptive mean shift clustering algorithm," in *Proc. SIAM Int. Conf. Data Mining*, 2014, pp. 794–802.
[18] Y. Ren, U. Kamath, C. Domeniconi, and G. Zhang, "Boosted mean shift clustering," in *Proc. ECML PKDD*, 2014, pp. 646–661.
[19] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 478–487.
[20] S. Huang, H. Wang, T. Li, T. Li, and Z. Xu, "Robust graph regularized nonnegative matrix factorization for clustering," *Data Mining Knowl. Discovery*, vol. 32, no. 2, pp. 483–503, 2018.
[21] D. Xie, Q. Gao, Q. Wang, and S. Xiao, "Multi-view spectral clustering via integrating global and local graphs," *IEEE Access*, vol. 7, pp. 31197–31206, 2019.
[22] L. Zheng, T. Li, and C. Ding, "Hierarchical ensemble clustering," in *Proc. IEEE 10th Int. Conf. Data Mining*, Dec. 2010, pp. 1199–1204.
[23] S. Mimaroglu and E. Aksehirli, "DICLENS: Divisive clustering ensemble with automatic cluster number," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 2, pp. 408–420, Mar./Apr. 2012.
[24] Y. Ren, C. Domeniconi, G. Zhang, and G. Yu, "Weighted-object ensemble clustering," in *Proc. Int. Conf. Data Mining*, Dec. 2013, pp. 627–636.
[25] Y. Ren, C. Domeniconi, G. Zhang, and G. Yu, "Weighted-object ensemble clustering: Methods and analysis," *Knowl. Inf. Syst.*, vol. 51, no. 2, pp. 661–689, 2017.
[26] H. F. Liu, "Spectral ensemble clustering," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 715–724.
[27] D. Huang, C.-D. Wang, and J.-H. Lai, "Locally weighted ensemble clustering," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1460–1473, May 2018.
[28] H. Liu, J. Wu, T. Liu, D. Tao, and Y. Fu, "Spectral ensemble clustering via weighted K-means: Theoretical and practical evidence," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 5, pp. 1129–1143, May 2017.
[29] Y.-P. Zhao, L. Chen, M. Gan, and C. L. P. Chen, "Multiple kernel fuzzy clustering with unsupervised random forests kernel and matrix-induced regularization," *IEEE Access*, vol. 7, pp. 3967–3979, 2018.
[30] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained K-means clustering with background knowledge," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 577–584.
[31] C. Ruiz, M. Spiliopoulou, and E. Menasalvas, "C-DBSCAN: Density-based clustering with constraints," in *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing* (Lecture Notes in Computer Science), vol. 4482. Berlin, Germany: Springer, 2007, pp. 216–223.
[32] H. Zeng and Y.-M. Cheung, "Semi-supervised maximum margin clustering with pairwise constraints," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 926–939, May 2012.
[33] Y. Ren, X. Hu, K. Shi, G. Yu, D. Yao, and Z. Xu, "Semi-supervised DenPeak clustering with pairwise constraints," in *Proc. 15th Pacific Rim Int. Conf. Artif. Intell.*, 2018, pp. 837–850.
[34] A. Arshad, S. Riaz, and L. Jiao, "Semi-supervised deep fuzzy C-mean clustering for imbalanced multi-class classification," *IEEE Access*, vol. 7, pp. 28100–28112, 2019.
[35] Z. Yu, P. Luo, J. You, H. S. Wong, H. Leung, S. Wu, and G. Han, "Incremental semi-supervised clustering ensemble for high dimensional data clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 701–714, Mar. 2016.
[36] N. Papernot, M. Abadi, Ú. Erlingsson, I. Goodfellow, K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," 2016, *arXiv:1610.05755*. [Online]. Available: https://arxiv.org/abs/1610.05755
[37] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–13.
[38] Z. Yu, Z. Kuang, J. Liu, H. Chen, J. Zhang, J. You, and G. Han, "Adaptive ensembling of semi-supervised clustering solutions," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 8, pp. 1577–1590, Aug. 2017.
[39] Z. Yu, P. Luo, J. Liu, H.-S. Wong, J. You, G. Han, and J. Zhang, "Semi-supervised ensemble clustering based on selected constraint projection," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 13, pp. 2394–2407, Dec. 2018.
[40] G. Karypis and V. Kumar, "A fast and high quality multilevel scheme for partitioning irregular graphs," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 359–392, Aug. 1999.
[41] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[42] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, pp. 2837–2854, Oct. 2010.

JIALIN TIAN is currently pursuing the bachelor's degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His current research interests include clustering, semi-supervised learning, and ensemble learning.



YAZHOU REN received the B.Sc. degree in information and computation science and the Ph.D. degree in computer science from the South China University of Technology, Guangzhou, China, in 2009 and 2014, respectively. He visited the Data Mining Laboratory, George Mason University, USA, from 2012 to 2014. He is currently an Associate Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. He has published more than 30 research articles. His current research interests include clustering, self-paced learning, and transfer learning.

XIANG CHENG received the bachelor's degree in electronic and electrical engineering from the University of Electronic Science and Technology of China (UESTC) and the University of Glasgow (UoG), in 2018. He is currently pursuing the Ph.D. degree in computer science with Virginia Tech. His research interests include deep learning, security and data mining, especially in interdisciplinary applications of security, and deep learning.