Taylor & Francis
Taylor & Francis Group

Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/taut20

Towards semi-supervised ensemble clustering using a new membership similarity measure

Wenjun Li, Ting Li & Musa Mojarad

To cite this article: Wenjun Li, Ting Li & Musa Mojarad (2023) Towards semi-supervised ensemble clustering using a new membership similarity measure, Automatika, 64:4, 764-771, DOI: 10.1080/00051144.2023.2217601

To link to this article: https://doi.org/10.1080/00051144.2023.2217601

AUTOMATIKA
2023, VOL. 64, NO. 4, 764–771
https://doi.org/10.1080/00051144.2023.2217601

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Towards semi-supervised ensemble clustering using a new membership similarity measure

Wenjun Li[a], Ting Li[b] and Musa Mojarad[c]

[a]School of Software and Service Outsourcing, Suzhou Vocational Institute of Industrial Technology, Suzhou, People's Republic of China; [b]Suzhou Blueprint Smart City Technology Co. Ltd, Suzhou, People's Republic of China; [c]Department of Computer Engineering, Firoozabad Branch, Islamic Azad University, Firoozabad, Iran

ABSTRACT

Hierarchical clustering is a common type of clustering in which the dataset is hierarchically divided and represented by a dendrogram. Agglomerative Hierarchical Clustering (AHC) is a common type of hierarchical clustering in which clusters are created bottom-up. In addition, semi-supervised clustering is a new method in the field of machine learning, where supervised and unsupervised learning are combined. Clustering performance is effectively improved by semi-supervised learning, as it uses a small amount of labelled data to aid unsupervised learning. Meanwhile, ensemble clustering by combining the results of several individual clustering methods can achieve better performance compared to each of the individual methods. Considering AHC with semi-supervised learning for ensemble clustering configuration has received less attention in the past literature. In order to achieve better clustering results, we propose a semi-supervised ensemble clustering framework developed based on AHC-based methods. Here, we develop a flexible weighting mechanism along with a new membership similarity measure that can establish compatibility between semi-supervised clustering methods. We evaluated the proposed method with several equivalent methods based on a wide variety of UCI datasets. Experimental results show the effectiveness of the proposed method from different aspects such as NMI, ARI and accuracy.

ARTICLE HISTORY
Received 31 December 2021
Accepted 20 May 2023

KEYWORDS
AHC; ensemble clustering; membership similarity measure; semi-supervised clustering

1. Introduction

Currently, there are different types of machine learning systems, which are classified into four general groups: supervised learning, unsupervised learning, Semi-Supervised Learning (SSL), and reinforcement learning [1,2]. Supervised learning includes data whose class labels are known and available in the learning phase. One of the common problems in this type of learning is the classification problem. Some of the most common classification algorithms are linear regression, logistic regression, k-nearest neighbours, support vector machine, neural networks, decision trees and random forests [3,4]. In unsupervised learning, the data class label is not available and the learning process seeks to assign the appropriate label to each data. One of the common problems in this type of learning is the clustering problem. In clustering, groups of similar objects should be identified. Some of the most common clustering algorithms are K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Hierarchical Cluster Analysis (HCA), FCM and c-means [5].

One of the successful approaches in recent years to improve clustering performance is ensemble clustering methods [6,7]. The main idea of learning in ensemble clustering is to combine the prediction results of different individual clustering models. Multiple clustering methods can create higher quality clusters by combining the output partitions of several basic models. In this regard, it can be expected that the use of ensemble clustering in the context of hierarchical clustering can provide a higher quality for creating the final partition [8]. According to the latest studies, the problem of ensemble hierarchical clustering has not received much attention so far. Hence, we draw inspiration from hierarchical clustering and SSL to develop an efficient ensemble clustering framework [9,10].

In this paper, a flexible weighting mechanism is developed to describe the consistency between semi-supervised clustering models used to generate base partitions. In general, the proposed algorithm consists of three main steps: creating primary clusters with different Agglomerative Hierarchical Clustering (AHC) methods [11], developing a new membership similarity measure to calculate the similarity between objects, and finally re-clustering the primary clusters to create final clusters. We generate primary clusters by four linkage-based AHC methods. The results are evaluated at the

CONTACT Wenjun Li liwenjun_sc@163.com School of Software and Service Outsourcing, Suzhou Vocational Institute of Industrial Technology, Suzhou, Jiangsu, 215000, People's Republic of China; Musa Mojarad musa.mojarad@iau.ac.ir Department of Computer Engineering, Firoozabad Branch, Islamic Azad University, Firoozabad, Iran

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

cluster and partition levels using a robustness measure to calculate the similarity between objects. We measure the weight of primary clusters based on their robustness. The primary clusters with the highest weight are selected for the final consensus to form the final partition. Here, the consensus function is developed based on the meta-clustering technique (i.e. re-clustering of the primary clusters). Finally, the final partition is created by assigning objects to meta-clusters with the highest similarity.

The main contribution of this paper is as follows:

- Configuration of a new membership similarity measure between objects inspired by the evaluation of clusters and partitions
- Development of a flexible weighting mechanism to generate consistent base partitions
- Improving the learning process in ensemble clustering using semi-supervised hierarchical clustering

The outline of the rest of the paper is as follows: Related works are reviewed in Section 2. General concepts related to clustering are given in Section 3. Section 4 explains the details of the proposed algorithm. Section 5 is related to the results of experiments and evaluations. Finally, Section 6 concludes the paper.

2. Related works

This section is a literature review to understand the problem of ensemble clustering and related concepts of semi-supervised framework [12–14]. A summary of the aforementioned studies is given in Table 1.

Zhang et al. [15] presented a Two-Stage approach for Semi-supervised Ensemble Clustering based on constraint weight (TSSEC). The authors propose some pairwise constraints to improve the clustering process: the supervised data is only used for the ensemble process, the final clusters are formed without considering the redundancy, and the influence of different clusters is ignored when forming the final clusters. To address these constraints, TSSEC can select appropriate clusters and consider cluster weights for the clustering task. Here, pairwise constraints are used to select clusters and cluster weights. TSSEC selects a subset of primary clusters based on the quality and diversity of the monitored data. The quality of selected clusters is calculated through unsupervised and supervised data. Finally, TSSEC uses a weighted correlation matrix to generate final clusters.

Yang et al. [16] proposed a semi-supervised consensus clustering approach using closed patterns. The authors developed their previous work on MultiCons multiple consensus clustering and presented the Semi-MultiCons approach. Semi-MultiCons does not depend on the number of clusters and creates final clusters based on different pairwise constraints. In addition, this approach can reduce the negative effects related to the integration of constraints in the clustering process.

Kadhim et al. [17] presented an ensemble clustering approach based on the Self-Directed Learning (SDL) framework. This approach can help the consensus function to achieve the highest evaluation in satisfying performance measurement. In general, SDL includes a combination of Predicting Test-set Labels (PTL) and Detecting Best Results (DBR). PTL combines clustering results sequentially to produce satisfactory results, where it helps to predict labels. Meanwhile, DBR can find the correct result when predicting several different results for the same model. In addition, the authors introduced new performance measurements for clustering validation, the most important of which is the Correction Ratio (CR).

Li et al. [18] proposed a new ensemble clustering algorithm for data with different scales. The authors introduce the Meta-Clustering Ensemble method based on Model Selection (MCEMS), which is a multi-step approach for data clustering. MCEMS tries to calculate the similarity between objects by considering several primitive clusters from different models. In addition, MCEMS is equipped with a clustering model selection technique considering quality and diversity.

3. Proposed algorithm

Ensemble clustering is proven to be an ideal alternative in terms of robustness and stability to an individual clustering algorithm [19]. The aim of this paper is to

Table 1. A summary of the reviewed studies.

Authors	Model name	Methodology	Strengths	Weakness
Zhang et al. [15]	TSSEC	Using pairwise constraints to select clusters and their weights	Calculation of the quality of selected clusters based on unsupervised and supervised data	Overhead of large searching spaces
Yang et al. [16]	Semi-MultiCons	Semi-supervised consensus clustering using closed patterns	No dependence on the number of clusters and creation of final clusters based on pair constraints	A small number of constraints is considered
Kadhim et al. [17]	SDL	Combination of PTL and DBR in ensemble clustering process	Introducing new performances measurement for clustering validation	Performance measurements are evaluated for small-scale data
Li et al. [18]	MCEMS	Meta-clustering ensemble method based on model selection	A new combinatorial mechanism for calculating similarity between objects	Slightly slow execution

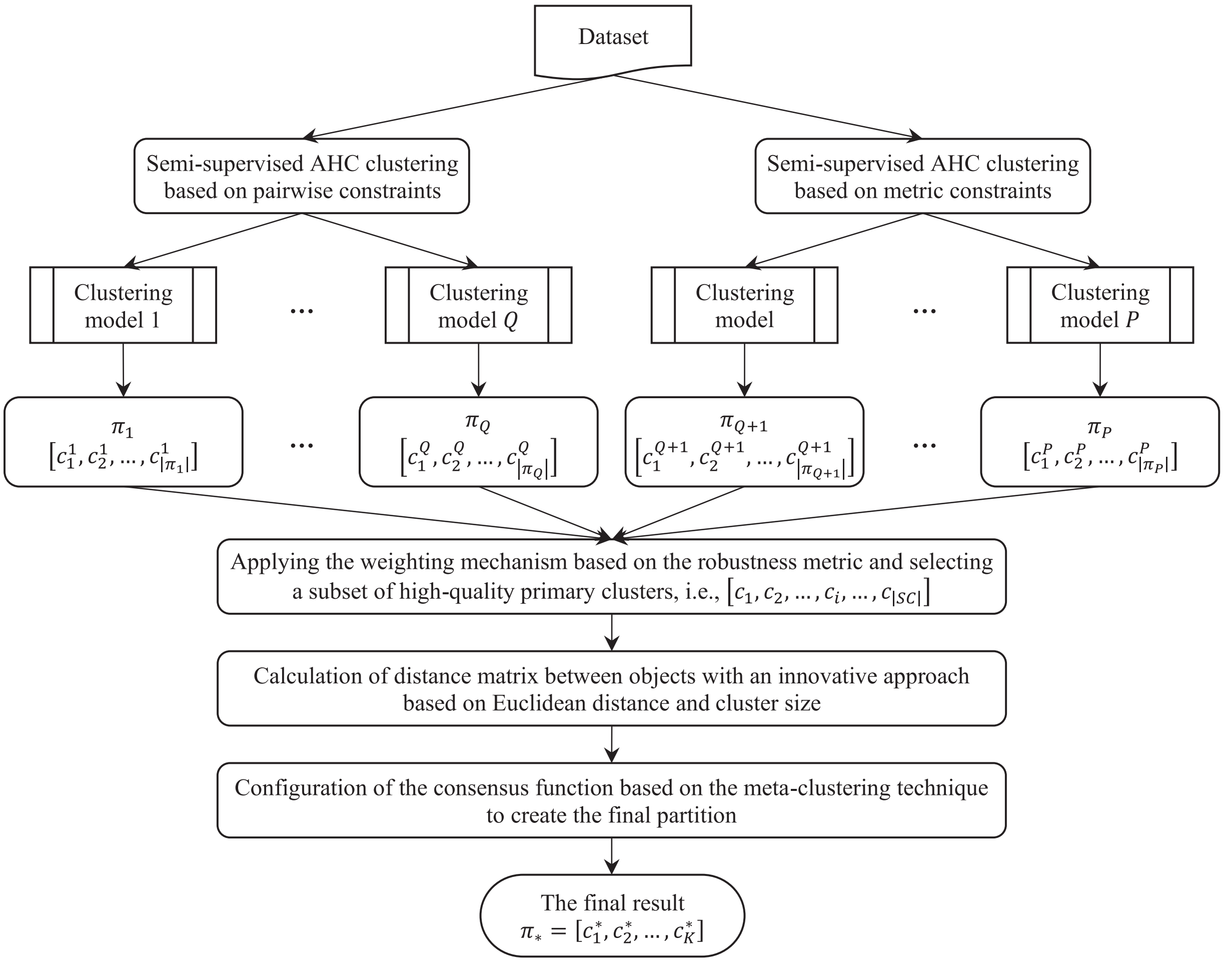
Dataset
Semi-supervised AHC clustering based on pairwise constraints
Semi-supervised AHC clustering based on metric constraints
Clustering model 1
…
Clustering model Q
Clustering model
…
Clustering model P
π_1 $[c_1^1, c_2^1, \dots, c_{|\pi_1|}^1]$
…
π_Q $[c_1^Q, c_2^Q, \dots, c_{|\pi_Q|}^Q]$
π_{Q+1} $[c_1^{Q+1}, c_2^{Q+1}, \dots, c_{|\pi_{Q+1}|}^{Q+1}]$
…
π_P $[c_1^P, c_2^P, \dots, c_{|\pi_P|}^P]$
Applying the weighting mechanism based on the robustness metric and selecting a subset of high-quality primary clusters, i.e., $[c_1, c_2, \dots, c_i, \dots, c_{|SC|}]$
Calculation of distance matrix between objects with an innovative approach based on Euclidean distance and cluster size
Configuration of the consensus function based on the meta-clustering technique to create the final partition
The final result $\pi_* = [c_1^*, c_2^*, \dots, c_K^*]$

Figure 1. Framework of the proposed clustering algorithm.

combine the advantages of SSL and ensemble clustering to improve clustering performance. Figure 1 describes the general framework of the proposed algorithm. First, the dataset is clustered by several semi-supervised AHC-based models. Two aspects are considered for applying SSL: information based on pairwise constraints and information based on metric constraints. This information can provide different aspects of the dataset with more flexibility for clustering. In both sections, we use four linkage-based AHC methods for clustering: single, centroid, average, and complete. Meanwhile, we present an innovative approach to measure the distance between objects, which is based on Euclidean distance and cluster size.

3.1. System model

Any clustering method can be applied to a given dataset and return a partition as output [20]. Let $X = \{x_1, x_2, \dots, x_N\}$ be a dataset with N objects. Here, $x_i \in X$ represents the i-th object of the dataset X. Let $x_i = [x_1^i, x_2^i, \dots, x_M^i]$ be the vector of M features corresponding to x_i. Let π be an individual clustering method and $\pi(x_i)$ is the label of the cluster belonging to x_i. In the ensemble clustering problem, X is clustered by a set of P individual methods. Let $\Pi = \{\pi_1, \pi_2, \dots, \pi_P\}$ be a set of P individual clustering methods where each method provides a partition as clustering output. Each partition contains several clusters that can be different according to the clustering methods used. Let $\pi_k = \left[c_1^k, c_2^k, \dots, c_{|\pi_k|}^k\right]$ be the primary clusters generated by the k-th member of the ensemble, where $|\pi_k|$ represents the partition size (number of clusters generated). The consensus function in ensemble clustering can provide the final partition by merging the generated partitions. Let $\pi_* = \Gamma\pi_1, \pi_2, \dots, \pi_P$ be a consensual consensus function applied to P generated partitions of Π. Here, Γ as a consensus function can produce the final partition π_*. Let $\pi_* = [c_1^*, c_2^*, \dots, c_K^*]$ be the final partition generated with K clusters obtained from the consensus of results in Π.

3.2. Semi-supervised AHC clustering based on pairwise constraints

Basically, constraint-based knowledge can lead to improved clustering performance, because it is easier to obtain than object labels. Pairwise constraints indicate whether a pair of objects can be included in a group or not [21]. In general, pairwise constraints include must-link and cannot-link. Let $ML = \{(x_i, x_j)\}$

denotes the must-link, where x_i and x_j can be grouped into a cluster. Also, let $CL = \{(x_i, x_j)\}$. denotes cannot-link, where x_i and x_j must be grouped into different clusters. Both must-link and cannot-link as pairwise constraints include properties of symmetry and transitivity. Let x_i, x_j and x_k be three objects of X. According to this, the properties of symmetry and transitivity in pair constraints are defined by Equations (1) and (2), respectively.

$$\begin{cases}(x_i, x_j) \in ML \rightarrow (x_j, x_i) \in ML\\(x_i, x_j) \in CL \rightarrow (x_j, x_i) \in CL\end{cases} \tag{1}$$

$$\begin{cases}(x_i, x_k) \in ML\&(x_k, x_j) \in ML \rightarrow (x_i, x_j) \in ML\\(x_i, x_k) \in CL\&(x_k, x_j) \in CL \rightarrow (x_i, x_j) \in CL\end{cases} \tag{2}$$

Let $d_{i,j} \in D$ be the distance between x_i and x_j in the distance matrix D. According to the definition of pair constraints, the distance matrix is defined. If $(x_i, x_j) \in ML$, then $d_{i,j} = 0$ and if $(x_i, x_j) \in CL$, then $d_{i,j} = \infty$. Meanwhile, let $s_{i,j} \in S$ be the similarity between x_i and x_j in the similarity matrix S. We define the similarity matrix by Equation (3).

$$s_{i,j} = exp\left(-\frac{x_i - {x_j}^2}{\sigma_i \sigma_j}\right) \tag{3}$$

where $||x_i - x_j||$ is equivalent to $d_{i,j}$, and σ_i and σ_j are the corresponding parameters for x_i and x_j, respectively. Here, σ_i is formulated by Equation (4).

$$\sigma_i = \frac{1}{N}\sum_{i=1}^{N} x_i - x_j \tag{4}$$

Finally, the clustering of the dataset X is done considering the similarity matrix S. Here, we use four linkage-based AHC clustering methods including single, centroid, average, and complete for clustering and creating partitions. All these methods provide clustering results by dendrogram. Each level of the dendrogram is considered as a partition. In this paper, Bayesian PAC learning [22] is used to select the appropriate level of the dendrogram and determine the appropriate partition. By determining the appropriate partition, the number of clusters (i.e. K) in each method is determined automatically.

3.3. Semi-supervised AHC clustering based on metric constraints

Huang et al. [23] proposed the large margin nearest cluster (LMNC) distance metric for semi-supervised clustering. LMNC is inspired by the Mahalanobis metric to realize the min–max principle. This principle states that robust clustering is achieved by minimizing the distances between objects in similar clusters and maximizing the distance between objects in different clusters [24]. Let $\{(x_i, y_j)\}_{i=1}^{N}$ be a dataset with N objects, where $x_i \in R^M$ refers to objects and $y_j \in \{1, 2, \ldots, K\}$ refers to class labels. Also, let $\mathcal{M}$ be a symmetric matrix of size $M \times M$. The distance square for each pair of objects x_i and x_j in the R^M space is formulated by Equation (5).

$$D(x_i, x_j) = (x_i - x_j)^T \mathcal{M}(x_i - x_j) \tag{5}$$

Basically, $\mathcal{M}$ is considered as a positive semi-definite matrix, where $\mathcal{M} \geq 0$. LMNC includes a cost function for learning the $\mathcal{M}$ matrix, as shown in Equation (6).

$$\begin{aligned}\varepsilon(L) = &\sum_{i,j} a_{i,j}(x_i - z_j)^T \mathcal{M}(x_i - z_j)\\&+ c\sum_{i,j} a_{i,j}(1 - a_{i,j})[1 + (x_i - z_j)^T \mathcal{M}(x_i - z_j)\\&- (x_i - z_j)^T \mathcal{M}(x_i - z_j)]_+\end{aligned} \tag{6}$$

where, $a_{i,j} \in \{0, 1\}$ represents the ordered weight with x_i and x_j. Here, $a_{i,j} = 1$ means that class label y_i and y_j are same for x_i and x_j respectively. Moreover, $c > 0$ is a positive constant, z_j is the centre of cluster j, and $[f]_+ = \max(f, 0)$ is the loss function. LMNC formulates the loss metric as an optimization problem to realize the min–max principle, as shown in Equation (7).

$$\begin{aligned}&\text{Min} \sum_{i,j} a_{i,j}(x_i - z_j)^T \mathcal{M}(x_i - z_j)\\&\quad + c\sum_{i,j,l} a_{i,j}(1 - a_{i,j})\xi_{i,j,l}\\&\text{s.t. (i) } \xi_{i,j,l} \geq 0, \text{(ii) } \mathcal{M} \geq 0, (x_i - z_l)^T \mathcal{M}(x_i - z_l)\\&\quad - (x_i - z_j)^T \mathcal{M}(x_i - z_j) \geq 1 - \xi_{i,j,l}\end{aligned} \tag{7}$$

where $\xi_{i,j,l}$ is used as a slack term to induce the loss function.

This optimization problem in LMNC is solved by gradient projection algorithm. Finally, the clustering of the dataset X is done considering the distance matrix D. In this section, four linkage-based AHC clustering methods including single, centroid, average, and complete are used for clustering. Similarly, Bayesian PAC learning technique is used to determine the appropriate level and optimal partition.

3.4. Weighing mechanism

In general, the robustness of a partition may be evaluated as weak, while it has one or more clusters of high quality. Therefore, it is not recommended to use all partitions as well as all primary clusters generated in the final consensus [5,25]. This may even lead to a decrease in the ensemble clustering performance and an increase in the computational complexity of the consensus function. Normalized Mutual Information (NMI) is a common performance metric for evaluating clustering. The evaluation in NMI is based on the diversity of labels in two partitions, as shown in Equation (8).

$$NMI(\pi_\alpha,\pi_\beta) = \frac{2\sum_{i=1}^{|\pi_\alpha|}\sum_{j=1}^{|\pi_\beta|} N_{ij}\log\left(\frac{N.N_{ij}}{N_{i\alpha}.N_{\beta j}}\right)}{\sum_{i=1}^{|\pi_\alpha|} N_{i\alpha}\log\left(\frac{N_{i\alpha}}{N}\right) + \sum_{j=1}^{|\pi_\beta|} N_{\beta j}\log\left(\frac{N_{\beta j}}{N}\right)} \qquad (8)$$

where, $\pi_\alpha = [c_1^\alpha, c_2^\alpha, \ldots, c_{|\pi_\alpha|}^\alpha]$ and $\pi_\beta = [c_1^\beta, c_2^\beta, \ldots, c_{|\pi_\beta|}^\beta]$ are two partitions, N represents the number of objects and N_{ij} represents the number of identical objects in c_i^α and c_j^β. Also, $N_{i\alpha}$ and $N_{i\alpha}$ represent the number of objects in c_i^α and c_j^β, respectively. Specifically, $NMI(\pi_\alpha, \pi_\beta) = 0$ indicates complete difference between π_α and π_β partitions, while $NMI(\pi_\alpha, \pi_\beta) = 1$ indicates complete similarity in these partitions.

Measuring the diversity by NMI between an output partition and reference partition can evaluate the quality of the clustering method. Therefore, the robustness of partitions created in Π can be measured by $Weight_{NMI}(\pi_\gamma) = NMI(\pi_\gamma, \pi_*)$. Here, we consider the robustness of a partition as its weight, where π_* represents the reference partition. With converting a cluster to a partition, NMI can be used to evaluate clusters. Let $Weight_{NMI}(c_i)$ be the weight of cluster c_i.

Let $AC = [c_1^1, c_2^1, \ldots, c_{|\pi_1|}^1, c_1^2, c_2^2, \ldots, c_{|\pi_2|}^2, \ldots, c_1^P, c_2^P, \ldots, c_{|\pi_P|}^P]$ be the set of all primary clusters of P partitions available. The goal is to select a subset of high-quality AC to participate in the consensus function. Let $SC = [c_1, c_2, \ldots, c_i, \ldots, c_{|SC|}]$ be the set of selected clusters from AC that participate in the final consensus. If $c_i \in SC$, then c_i satisfies the predefined threshold. We define this threshold based on $Weight_{NMI}(c_i) \geq \theta$, where θ is a fixed parameter to determine the merit of the clusters. Experimentally, θ is set to 0.35.

4. Results and discussion

In this section, we evaluate the proposed algorithm and its results. All experiments were performed by the MATLAB 2021a simulator on a desktop with Intel® Core™ i7-2600 Processor (8M Cache, up to 3.80 GHz), 32 GB of RAM DDR4 and 64-bit Windows 10. We use various evaluation metrics to demonstrate the superiority of the proposed algorithm, for example, NMI, Adjusted Rand Index (ARI), accuracy and running time.

4.1. Datasets

In order to evaluate the proposed algorithm in comparison with other existing clustering methods, several different datasets from the UCI machine learning repository have been used. Table 2 shows the characteristics of these datasets.

Table 2. Details of datasets used in the experiments.

Dataset name	Number of objects	Number of features	Number of classes
Iris	150	4	3
Wine	178	13	3
BNG Spect	1000000	23	2
Voice_9	428	10	9
Road Safety	363243	67	2
Glass	214	9	6
BNG Vote	131071	17	2
Thyroid	215	5	3
Secom	1567	590	2
Waveform	5000	21	3

4.2. Discussion and comparisons

This section is related to the evaluation and validation of the proposed algorithm in terms of different performance metrics. We compare the proposed algorithm based on NMI, ARI, accuracy and running time with some equivalent methods such as TSSEC [15], Semi-MultiCons [16], SDL [17] and MCEMS [18].

The accuracy of the proposed algorithm in clustering compared to the existing methods is shown in Figure 2. The proposed algorithm and each of the methods are compared in a subplot. The results show the superiority of the proposed algorithm in most of the datasets. The proposed algorithm outperforms TSSEC and MCEMS in all datasets. The average superiority over all datasets is reported as 13.41% and 15.18%, respectively. Compared to SDL, the proposed algorithm has absolute superiority in all datasets except Voice_9 and Secom. The accuracy results show that on average the proposed algorithm is more than 6.5% superior to the SDL method. As illustrated, the results of the proposed algorithm are competitive compared to Semi-MultiCons. However, the proposed algorithm provides an average of 3.34% better accuracy than this method.

Table 3 shows the average performance calculated by the ARI metric through the standard deviation. These results for the NMI metric are reported in Table 4. Meanwhile, the runtime for each method is reported in Table 5. The bold results in these tables represent the best values for each method. The results clearly prove the better performance of the proposed algorithm. As illustrated, the results of the proposed algorithm are better compared to existing methods on large-scale datasets. This is clearly evident when looking at the results associated with the BNG Spect and BNG Vote datasets. Compared to TSSEC, MCEMS, SDL and Semi-MultiCons, the proposed algorithm is superior in ARI metric by 20.49%, 12.68%, 8.75% and 1.69%, respectively. This superiority for NMI metric is reported as 7.97%, 11.27%, 4.39% and 1.76%, respectively. In terms of runtime, the proposed algorithm has the least complexity on average.

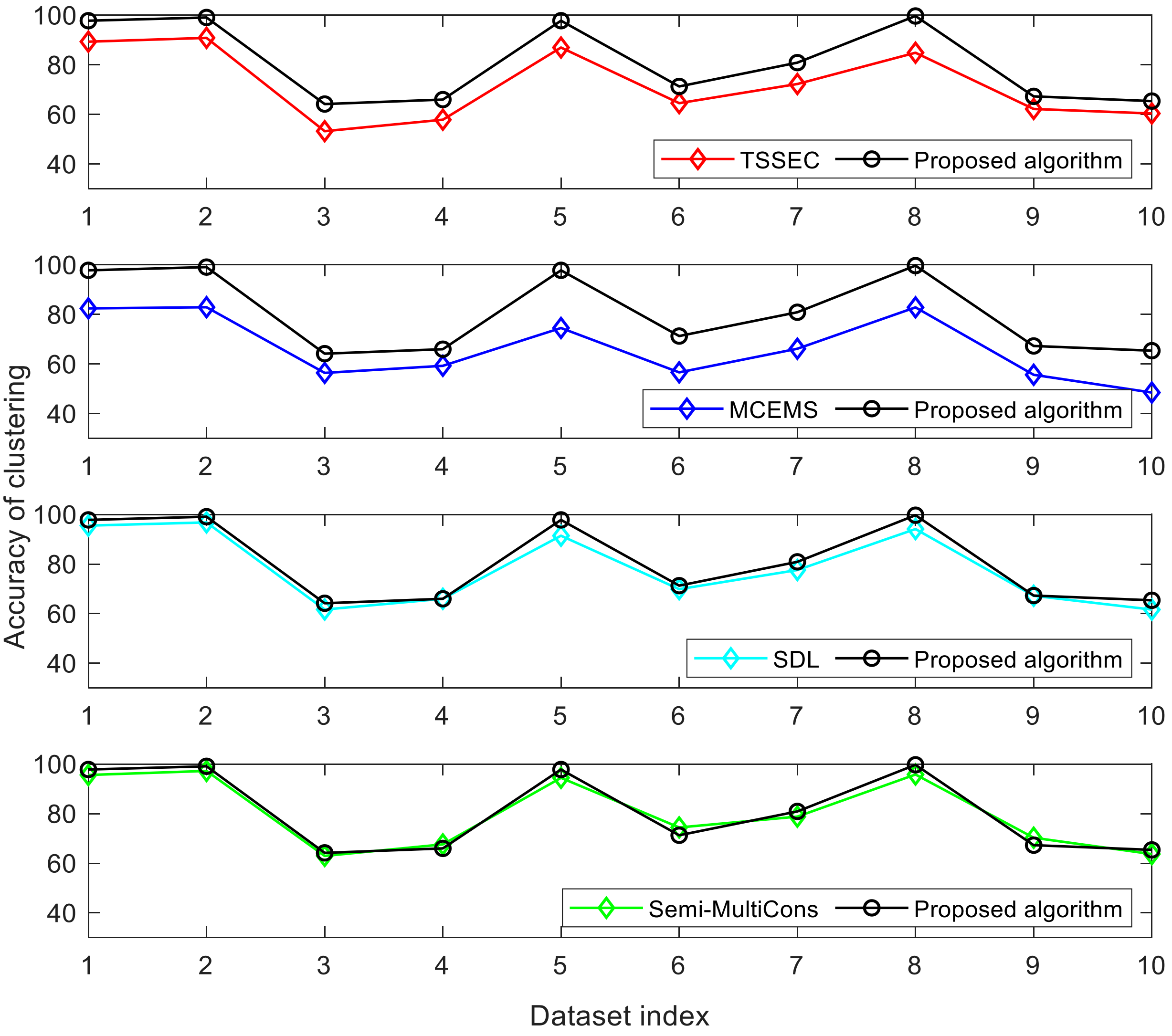

Figure 2. Comparison of different methods based on clustering accuracy.

Table 3. ARI results for different methods.

Dataset name	TSSEC	MCEMS	SDL	Semi-MultiCons	Proposed algorithm
Iris	0.8869 ± 0.024	0.8838 ± 0.057	0.8992 ± 0.065	**0.9084 ± 0.000**	0.8983 ± 0.002
Wine	0.2574 ± 0.029	0.2841 ± 0.031	0.2913 ± 0.070	0.3273 ± 0.054	**0.3447 ± 0.017**
BNG Spect	0.2367 ± 0.063	0.3861 ± 0.023	0.4951 ± 0.007	0.5121 ± 0.068	**0.5454 ± 0.060**
Voice_9	0.3422 ± 0.079	0.4298 ± 0.038	0.4123 ± 0.004	0.4327 ± 0.031	**0.4596 ± 0.047**
Road Safety	0.6748 ± 0.025	0.6957 ± 0.035	0.5247 ± 0.018	**0.8005 ± 0.007**	0.7461 ± 0.027
Glass	0.5220 ± 0.063	0.5309 ± 0.075	**0.6155 ± 0.060**	0.5820 ± 0.054	0.5681 ± 0.031
BNG Vote	0.5639 ± 0.047	0.5945 ± 0.022	0.6831 ± 0.072	0.7129 ± 0.014	**0.7317 ± 0.026**
Thyroid	0.7369 ± 0.058	0.7109 ± 0.025	0.7511 ± 0.017	0.7487 ± 0.013	**0.7902 ± 0.019**
Secom	0.7918 ± 0.064	0.7576 ± 0.072	0.8513 ± 0.003	0.8742 ± 0.043	**0.9081 ± 0.015**
Waveform	0.3230 ± 0.017	0.4319 ± 0.048	0.3880 ± 0.002	0.4233 ± 0.037	**0.4366 ± 0.072**

Table 4. NMI results for different methods.

Dataset name	TSSEC	MCEMS	SDL	Semi-MultiCons	Proposed algorithm
Iris	0.7854 ± 0.022	0.7865 ± 0.016	0.7973 ± 0.074	0.7931 ± 0.073	**0.7988 ± 0.003**
Wine	0.4509 ± 0.060	0.3235 ± 0.022	0.4596 ± 0.073	**0.5004 ± 0.051**	0.5002 ± 0.043
BNG Spect	0.3317 ± 0.033	0.3292 ± 0.000	0.3317 ± 0.013	0.3457 ± 0.051	**0.3631 ± 0.075**
Voice_9	0.4235 ± 0.008	0.4290 ± 0.075	0.4349 ± 0.004	0.4356 ± 0.056	**0.4724 ± 0.006**
Road Safety	0.7724 ± 0.041	0.7625 ± 0.047	**0.8089 ± 0.006**	0.7756 ± 0.057	0.7854 ± 0.040
Glass	0.5219 ± 0.031	0.5158 ± 0.044	0.5307 ± 0.052	0.5657 ± 0.036	**0.5733 ± 0.058**
BNG Vote	0.5878 ± 0.062	0.5908 ± 0.041	0.5998 ± 0.065	0.6376 ± 0.056	**0.6611 ± 0.033**
Thyroid	0.7371 ± 0.052	0.7289 ± 0.064	0.7489 ± 0.030	**0.7693 ± 0.073**	0.7678 ± 0.030
Secom	0.7813 ± 0.064	0.7518 ± 0.033	0.8561 ± 0.007	0.8704 ± 0.073	**0.8860 ± 0.029**
Waveform	0.4187 ± 0.037	0.4204 ± 0.022	0.4422 ± 0.069	**0.4723 ± 0.063**	0.4660 ± 0.030

Table 5. Running time (s) results for different methods.

Dataset name	TSSEC	MCEMS	SDL	Semi-MultiCons	Proposed algorithm
Iris	4.80 ± 2.21	4.74 ± 1.06	5.23 ± 1.64	**4.36 ± 1.33**	4.62 ± 1.43
Wine	5.24 ± 2.47	4.85 ± 1.57	6.37 ± 2.13	4.11 ± 1.44	4.07 ± 1.01
BNG Spect	905.28 ± 41.16	885.67 ± 25.77	915.21 ± 36.20	865.34 ± 19.99	**846.45 ± 22.94**
Voice_9	228.68 ± 14.51	225.14 ± 9.03	**216.00 ± 10.74**	241.35 ± 6.37	218.07 ± 7.87
Road Safety	834.32 ± 40.40	828.37 ± 34.23	883.28 ± 31.23	**785.36 ± 29.36**	816.46 ± 34.44
Glass	157.93 ± 11.89	157.42 ± 7.63	169.83 ± 8.56	156.40 ± 5.02	**146.02 ± 6.14**
BNG Vote	570.05 ± 38.90	560.47 ± 34.41	577.54 ± 28.37	562.56 ± 23.70	**541.32 ± 18.66**
Thyroid	34.48 ± 6.16	32.70 ± 6.39	34.85 ± 4.29	34.10 ± 2.33	**29.16 ± 3.02**
Secom	91.89 ± 7.90	90.73 ± 7.42	90.41 ± 5.32	93.36 ± 5.60	**88.34 ± 4.26**
Waveform	429.29 ± 17.47	423.10 ± 15.11	441.76 ± 18.02	**410.73 ± 14.06**	416.81 ± 13.84

5. Conclusions

In this paper, we developed AHC-based ensemble clustering inspired by SSL. Here, we develop a flexible weighting mechanism that can describe the consistency between semi-supervised clustering methods used to generate base partitions. Also, we presented a new membership similarity measure to calculate the similarity between objects that uses the results from evaluating clusters and partitions simultaneously. Evaluations on some datasets from the UCI repository show that the proposed algorithm is significantly superior compared to equivalent methods. This superiority exists in many performance metrics such as NMI, ARI and accuracy. For future work, we develop the proposed algorithm for modelling to avoid reassembling the entire dataset in each run.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Natural Science Foundation of Jiangsu Province under Grant BK20180209. Natural Science Foundation of the Jiangsu Higher Education Institutions of China under Grant 18KJD520004. Research Funds of Suzhou Vocational Institute of Industrial Technology 2019kyqd018.

References

[1] Casas P, Mazel J, Owezarski P. MINETRAC: mining flows for unsupervised analysis & semi-supervised classification. In: 2011 23rd International Teletraffic Congress (ITC). IEEE; 2011, September. p. 87–94.

[2] Adıyeke E, Baydoğan MG. Semi-supervised extensions of multi-task tree ensembles. Pattern Recognit. 2022;123:108393.

[3] Yang T, Pasquier N, Precioso F. Ensemble clustering based semi-supervised learning for revenue accounting workflow management. In: DATA. 2020, July. p. 283–293.

[4] He Y, Chang XH, Wang H, et al. Command-filtered adaptive fuzzy control for switched MIMO nonlinear systems with unknown dead zones and full state constraints. Int J Fuzzy. 2022. DOI: 10.1007/s40815-022-01384-y.

[5] Zhou J, Zhu SF, Huang X, et al. Enhancing time series clustering by incorporating multiple distance measures with semi-supervised learning. J Comput Sci Technol. 2015;30(4):859–873.

[6] Mahmood A, Li T, Yang Y, et al. Semi-supervised evolutionary ensembles for web video categorization. Knowl Based Syst. 2015;76:53–66.

[7] Cao C, Wang J, Kwok D, et al. webTWAS: a resource for disease candidate susceptibility genes identified by transcriptome-wide association study. Nucleic Acids Res. 2022;50(D1):D1123–D1130.

[8] Fang Q, Liu X, Zeng K, et al. Centrifuge modelling of tunnelling below existing twin tunnels with different types of support. Underground Space. 2022;7(6): 1125–1138.

[9] Li P, Yang M, Wu Q. Confidence interval based distributionally robust real-time economic dispatch approach considering wind power accommodation risk. IEEE Trans Sustain Energy. 2021;12(1):58–69.

[10] Tan J, Liu L, Li F, et al. Screening of endocrine disrupting potential of surface waters via an affinity-based biosensor in a rural community in the Yellow River basin, China. Environ Sci Technol. 2022;56(20): 14350–14360.

[11] Cheng F, Liang H, Wang H, et al. Adaptive neural self-triggered bipartite fault-tolerant control for nonlinear MASs With dead-zone constraints. IEEE Trans Autom Sci Eng. 2022. DOI: 10.1109/TASE.2022.3184022.

[12] Zhang H, Wang H, Niu B, et al. Sliding-mode surface-based adaptive actor-critic optimal control for switched nonlinear systems with average dwell time. Inf Sci (Ny). 2021;580:756–774.

[13] Tang F, Niu B, Zong G, et al. Periodic event-triggered adaptive tracking control design for nonlinear discrete-time systems via reinforcement learning. Neural Netw. 2022;154:43–55.

[14] Li T, Rezaeipanah A, El Din EMT. An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement. J King Saud Univ Comput Inform Sci. 2022;34(6):3828–3842.

[15] Zhang D, Yang Y, Qiu H. Two-stage semi-supervised clustering ensemble framework based on constraint weight. Int J Mach Learn Cybernet. 2022;14:567–586.

[16] Yang T, Pasquier N, Precioso F. Semi-supervised consensus clustering based on closed patterns. Knowl Based Syst. 2022;235:107599.

[17] Kadhim MR, Zhou G, Tian W. A novel self-directed learning framework for cluster ensemble. J King Saud Univ Comput Inform Sci. 2022;34(10):7841–7855.

[18] Li Y, Niu B, Zong G, et al. Command filter-based adaptive neural finite-time control for stochastic nonlinear systems with time-varying full-state constraints and asymmetric input saturation. Int J Syst Sci. 2022;53(1):199–221.

[19] Zhang H, Zhao X, Zong G, et al. Fully distributed consensus of switched heterogeneous nonlinear multi-agent systems with Bouc-Wen hysteresis input. IEEE Trans Netw Sci Eng. 2022;9(6):4198–4208.

[20] Zhang H, Zou Q, Ju Y, et al. Distance-based support vector machine to predict DNA N6-methyladenine modification. Curr Bioinf. 2022;17(5):473–482.

[21] Liu Z, Zheng Z, Sudhoff SD, et al. Reduction of common-mode voltage in multiphase two-level inverters using SPWM with phase-shifted carriers. IEEE Trans Power Electron. 2016;31(9):6631–6645.

[22] Seldin Y, Tishby N. PAC-Bayesian Analysis of Co-clustering and beyond. J Mach Learn Res. 2010;11(12): 3595–3646.

[23] Huang M, Chen Y, Liu J, et al. A large margin nearest cluster metric based semisupervised clustering algorithm for brain fibers. In: The 2014 5th international conference on game theory for networks. IEEE; 2014, November. p. 1–5.

[24] Si Z, Yang M, Yu Y, et al. Photovoltaic power forecast based on satellite images considering effects of solar position. Appl Energy. 2021;302:117514.

[25] Wang M, Yang M, Fang Z, et al. A practical feeder planning model for urban distribution system. IEEE Trans Power Syst. 2022. DOI: 10.1109/TPWRS.2022.3170933.