Wised Semi-Supervised Cluster Ensemble Selection: A New Framework for Selecting and Combining Multiple Partitions Based on Prior Knowledge

Fozieh Asghari Paeenroodposhti, Saber Nourian, Muhammad Yousefnezhad

1) Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran
2) Department of Electrical Engineering, Sari Branch, Islamic Azad University, Sari, Iran
3) College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China

fozasghari@yahoo.com; snourian@elec.iust.ac.ir; myousefnezhad@nuaa.edu.cn

Received: 2016/03/08; Accepted: 2016/07/09

Abstract

The Wisdom of Crowds, an innovative theory described in social science, claims that the aggregate decisions made by a group will often be better than those of its individual members if the four fundamental criteria of this theory are satisfied. This theory used for in clustering problems. Previous research showed that this theory can significantly increase the stability and performance of learning problems. As a solution, this paper proposes a new methodology of using WOC theory for evaluating and selecting basic result partitions in semi-supervised clustering problems. This paper introduces new technique for reducing the data dimensions based on supervision information, a new semi-supervised clustering algorithm based on k-means for generating basic results, a new strategy for evaluating and selecting basic results based on feedback mechanism, a new metric for evaluating diversity of basic results. The results demonstrate the efficiency of proposed method’s aggregate decision-making compared to other algorithms.

Keywords: Semi-Supervised Learning, Cluster Ensemble Selection, Wisdom of Crowds, Pairwise Constraints, Constraint Projection

1. Introduction

Clustering one of the main tasks in data mining, discovers meaningful patterns in the non-labeled data sets. In other words, it is the process of grouping data points into clusters so that members of the same cluster are more similar to each other than to members of other clusters, and has been a very active area in machine learning, pattern recognition, data mining, and etc. for many years [1], [2]. Generally, different algorithms provide predictions with different accuracy rates. However, selecting the best model is not necessarily the ideal choice because potentially valuable information may be wasted by discarding the results of less-successful models [3], [4], [5], [6], [7]. This leads to the concept of combining, where the outputs (individual predictions) of several models are pooled to make a better decision (collective prediction) [8], [9].

Research in the Clustering Combination field has shown that these pooled outputs have more strength, novelty, stability, and flexibility than the results provided by individual algorithms [3], [8].

Recently, semi-supervised clustering [11], [12], [13] which uses prior supervision information to aid the clustering process, has received a considerable amount of
attention. Prior information can be available in diverse forms [14], [15] such as labeled data, known data groups or associations, pairwise constraints, and etc. This paper focuses on pairwise constraints, i.e. pairs of instances known as belonging to the same class (must-link constraints) or different classes (cannot-link constraints). Pairwise constraints arise naturally in many real tasks and have been widely used in semi-supervised clustering. While pairwise constraints have potential to improve clustering accuracy, in practice, they often result in highly unstable clustering performance [17], [18]. The reason is that the composition and cardinality of constraint sets can significantly affect the performance of semi-supervised clustering. Recently, some researchers realize this important issue and propose several ways for selecting more informative constraints for semi-supervised clustering [18], [19], [20]. However, to the best of our knowledge, choosing appropriate constraint sets for specific algorithms and tasks is still very challenging.

In the social science arena, there is a corresponding research field known as the Wisdom of Crowds (WOC), simply claiming that the WOC is the phenomenon whereby decisions made by aggregating the information of groups usually have better results than those made by any single group member.

In recent years, the WOC used for proposing new frameworks in classification and clustering problems [7], [22]. Previous researches showed that this theory can significantly increase the stability and performance of learning problems. First time in semi-supervised learning, this paper proposes a new methodology of using pairwise constraints for semi-supervised Cluster Ensemble Selection based on the Wisdom of Crowds theory, which is called Wised Semi-Supervised Cluster Ensemble Selection (WSCES). Experimental results on a various number of real world datasets, show that in comparison with other clustering methods, WSCES improves the final results more efficiently.

The rest of this paper is organized as follows: In Section 2, this study first briefly reviews some related works on cluster ensemble selection, semi-supervised clustering, wisdom of crowds, and its application in supervised and unsupervised learning. Then it introduces the proposed Wised Semi-supervised Clustering Ensemble Selection framework in Section 3. Experimental results are reported in Section 4; and finally this paper presents discussion and point out some future works in Section 5.

2. Literature Review

2.1 Cluster Ensemble Selection

Clustering is the art of grouping data points into clusters so that members of the same cluster are more similar to each other than to members of other clusters. Generally, different clustering algorithms, which is called basic algorithms in this paper, provide predictions with different accuracy rates. In practice, basic algorithms cannot provide stable and accurate results, which is called basic results in this paper, in real world applications. For solving this problem, cluster ensemble proved that better final results can be generated by combining basic results instead of only choosing the best one. Generally, a cluster ensemble has two important steps [3], [4], [5], [6], [7], [8]:

1. Generating different results from primary clustering methods using different algorithms and changing the number of their partitions. This step is called generating diversity or variety.
2. Combining the primary results and generating the final ensemble. This step is performed by consensus functions (aggregating mechanism).

In Cluster Ensemble Selection [3] was proposed this idea that "All achieved basic results are not suitable for cooperating to generate the final result". Instead of combing all achieved basic results, we can combine a selected group of best basic results according to consensus metric(s) from ensemble committee in order to improve the accuracy of final results [3], [4], [5], [6], [7]. It is clear that an ensemble with a set of identical models cannot provide any advantages. Thus, the aim is to combine models that predict different outcomes, and there are four parameters dataset, clustering algorithms, evaluation metrics, and combine methods that can be changed to achieve this goal.

Strehl and Ghosh proposed the MI for measuring the consistency of data partitions [8]; Fred and Jain proposed NMI, which is independent of cluster size. This metric can be used to evaluate clusters and the partitions in many applications [23]. For instance, Zhong and Ghosh used NMI for evaluating clusters in document clustering [24], Kandylas et al. used it for community knowledge analysis [25], and Long et al. used it for evaluating graph clustering [26]. Fern and Lin developed a method which effectively uses a selection of basic partitions to participate in the ensemble, and consequently in the final decision. They also used the SNMI and Pairwise NMI as quality and diversity metrics between partitions, respectively [3]. Azimi and Fern used cluster ensemble selection to avoid consensus partitions which are excessively different from the base partitions they result from. They demonstrated that their method can result in partitions with enhanced SNMI [27]. Alizadeh et al. have explored the disadvantages of NMI as a symmetric criterion. They used the APMM and MAX metrics to measure diversity and stability, respectively, and suggested a new method for building a co-association matrix from a subset of base cluster results [4], [5], [6]. This paper introduces Uniformity for diversity measurement, which works based on the APMM metric in Section 3.3.1.

2.2 Semi-Supervised Clustering

Semi-supervised clustering uses supervision information to aid the clustering process. Supervision information can be in the form of class labels [11], pairwise constraints [16], [28], [29] etc. Pairwise constraints are frequently used in semi-supervised clustering because in many applications considering pairwise constraints is more practical than trying to obtain class labels [15], [30].

The main approaches for pairwise constraints based semi-supervised clustering fall into two general categories, i.e. constraint-based methods and metric-based methods. In constraint-based methods, the pairwise constraints are often used to enforce constraints during the clustering process. For example, Wagstaff et al. [17] proposed the constrained k-means method by introducing such pairwise constraints into standard k-means algorithm. Chang and Yeung [31] proposed a locally linear metric based on pairwise constraints. Yan and Domeniconi [32] proposed the subspace metric ensemble method for metric learning in high dimensional space. Recently, Tang et al. [33] proposed a feature projection method from pairwise constraints. Zhang et al. [34] proposed semi-supervised dimensionality reduction to project data into low-dimensional space using both pairwise constraints and unlabeled data.

All above mentioned algorithms show that pairwise constraints have potential to improve clustering accuracy, but different constraint sets usually result in highly unstable clustering performance. In other words, the values of pairwise constraints are not
identically important. There are some good constraint sets, which are beneficial for the clustering problem concerned. Also, there exist some bad constraint sets, which are helpless or even harmful for the clustering problem concerned. So a natural and key question is, given a set of pairwise constraints, can we decide in advance whether good or bad it is? Or equivalently, can we select the most appropriate (good) subset of constraints from a possible large constraint set?

Recently, some researchers begin to address that important issue and have proposed several ways for selecting more informative constraints for semi-supervised clustering. Wagstaff et al. [17] proposed the inconsistency and incoherence measures to evaluate the usefulness of constraints. Greene and Cunningham [18] proposed an ensemble approach to identify informative constraints for semi-supervised clustering. Davidson and Ravi [19] presented some intractability results regarding on clustering with constraints. In another related work, Zhang and Yan [35] analyzed the value of pairwise constraints in classification problem. Anand et al. proposed a semi-supervised framework for kernel mean shift clustering (SKMS) that uses only pairwise constraints to guide the clustering procedure. They used the initial kernel matrix by minimizing a LogDet divergence-based objective function for first mapped to a high-dimensional kernel space where the constraints are imposed by a linear transformation of the mapped points [36]. Xiong et al. proposed Neighborhood-based Framework (NBF) method. This method builds on the concept of neighborhood, where neighborhoods contain "labeled examples" of different clusters according to the pairwise constraints. Also, it expands the neighborhoods by selecting informative points and querying their relationship with the neighborhoods [37].

2.3 The Wisdom of Crowds

The Wisdom of Crowds [21] presents numerous case studies, primarily in economics and psychology, to illustrate how the prediction performance of a crowd is better than that of its individual members. The book relates to diverse collections of independent individuals, rather than crowd psychology as traditionally understood. Its central thesis, that a diverse collection of individuals making independent decisions is likely to make certain types of decisions and predictions better than individuals or even experts, draws many parallels with statistical sampling, but there is little overt discussion of statistics in the book. Mackey [7], [21], [22], [38] mentions that not every crowd is wise. These key criteria separate wise crowds from irrational ones: [21]

- Diversity of opinion: Each person has private information, even if it is only an eccentric interpretation of the known facts.
- Independence: People's opinions are not determined by the opinions of those around them.
- Decentralization: People are able to specialize and draw on local knowledge.
- Aggregation: Some mechanism exists for turning private judgments into a collective decision.

It is important to note that, under some conditions, the cooperation of the crowd will fail because of the consciousness of its members about each other's opinion. This will lead them to conform rather than think differently. Although each member of the crowd may attain greater knowledge and intelligence by this effect, definitely the whole crowd as a whole will become trapped into less unwise [7], [21], [38].
In supervised learning, Steyvers et al. [39] used WOC for recollecting order information, and Miller et al. [40] proposed an approach to the rank ordering problem. Finally, Baker and Ellison [22] proposed a method for using the WOC in ensembles and modules in environmental modeling.

In unsupervised learning, Yousefnezhad et al. used the definition of Independency on WOC theory for estimating the effects of basic parameters of basic clustering algorithms [10], for instance first initial center points in k-means algorithm, on the final result of cluster ensemble selection. They proposed Basic Parameters Independency (BPI) function for this estimation. Alizadeh et al. proposed Wisdom of Crowds Cluster Ensemble (WOCCE). They redefined WOC criteria for cluster ensemble selection literature [7]. This paper uses some of these definitions and changes them based on semi-supervised learning literature.

### 3. Proposed Method

Based on Surowiecki outlines [7], [21], the conditions for a crowd to be wise are: diversity, independence, decentralization, and aggregation method. To map the WOC to a semi-supervised cluster ensemble selection, this paper should restate the wise crowd requirements for the corresponding field. This section discusses these preconditions in detail for the area of semi-supervised clustering.

It seems that the best matching between individuals and their opinions in WOC is base clustering algorithms and partitions, respectively, in the context of cluster ensembles. The goal of WSCES, as illustrated in Fig. 1, is to construct a wise crowd in the primary partition via a recursive procedure. In the diversity stage, basic results are evaluated by Uniformity metric and the evaluated result is added to ensemble committee only if it has an acceptable diversity degree, which is determined by Threshold Value Estimation (TVE) method. The above three stages are repeated until the number of ensemble committee members reach to enough amount.

Then, the final result is created by using the members of ensemble committee and aggregation method in the last stage. The rest of this section each of components in the framework will be described.

#### 3.1 Independency

As mentioned before, semi-supervised methods use supervision information to aim clustering algorithm for solving problems. In this paper, the form of supervision information is pairwise constraints, which contain must- link and cannot-link instances.
Based on independency definition, "People’s opinions are not determined by the opinions of those around them", the options must be independence. This method uses basic algorithms and basic results instead of people and options [7], [22]. So, each basic clustering algorithm must generate clustering basic result, independently. In other words, the knowledge of solving clustering problem for each clustering algorithm must be independence. In fact, this is one of famous methods for generating diversify basic results in cluster ensemble selection approach [3], [4], [5], [6], [7], [10], [23]. The main reason of independency criterion in the proposed method is generating a set of parameters, which is called Independency Parameter (IP), based on pairwise constraints for aiming basic clustering algorithms to generate basic results, independently. IP can reduce the size of original data set, and also can reform data set to make more diversify results.

3.1.1 Random Selection

The proposed method categorize constraints into global constraints (or common knowledge) and local constraints (or private knowledge). The former is shared by all the individuals, while the latter is private for current individual only. This method randomly choose half constraints in a constraint set as global constraints and divide the remaining half constraints into disjoint parts, each of which corresponds to a set of local
constraints. A constraint subset for individual clustering consists of all global constraints plus respective set of local constraints.

### 3.1.2 Constraint Projection

Given a set of high-dimensional data examples \( X = \{x_1, x_2, ..., x_n\} \), and the corresponding pairwise must-link constraint set \( M = \{(x_i, x_j) \mid x_i \text{ and } x_j \text{ belong to the same class}\} \) and pairwise cannot-link constraint set \( C = \{(x_i, x_j) \mid x_i \text{ and } x_j \text{ belong to different classes}\} \), Constraint Projection seeks a set of projective vectors \( W = [w_1, w_2, ..., w_d] \), such that the pairwise constraints in \( C \) and \( M \) are most faithfully preserved in the transformed low-dimensional representations \( y_i = W^T x_i \). That is, examples involved by \( M \) should be close while examples involved by \( C \) should be far in the low-dimensional space. Define the objective function as maximizing \((W)\) with respect to \((w.r.t.) \ W^T W = I\), where:

\[
J(W) = \frac{1}{2n_c} \sum_{(i, x_i) \in C} y_i - y_j - \frac{\gamma}{2n_M} \sum_{(i, x_i), (j, x_j) \in M} y_i - y_j - \frac{1}{2n_c} \sum_{(i, x_i) \in C} W^T x_i
\]

\[
-W^T x_i^2 - W^T x_j^2
\]

Where \( n_c \) and \( n_M \) are the cardinality of the cannot-link constraints set \( C \) and the must-link constraints set \( M \) respectively, and \( \gamma \) is a scaling coefficient.

The intuition behind (1) is to let the average distance in the low-dimensional space between examples involved by the cannot-link set \( C \) as large as possible, while distances between examples involved by the must-link set \( M \) as small as possible. Since the distance between examples in the same class is typically smaller than that in different classes, a scaling parameter \( \gamma \) is added to balance the contributions of the two terms in (1) and its value can be estimated as follows:

\[
\gamma = \frac{1}{n_c} \sum_{(i, x_i) \in C} x_i - x_j^2
\]

\[
\frac{1}{n_M} \sum_{(i, x_i), (j, x_j) \in M} x_i - x_j^2
\]

We can also reformulate the objective function in (1) in a more convenient way as follows:

\[
J(W) = trace \left( W^T \left[ S_c - \gamma S_M \right] W \right)
\]

Where \( S_C \) and \( S_M \) are respectively defined as:

\[
S_C = \frac{1}{2n_c} \sum_{(i, x_i), (j, x_j) \in C} (x_i - x_j)(x_i - x_j)^T
\]

\[
S_M = \frac{1}{2n_M} \sum_{(i, x_i), (j, x_j) \in M} (x_i - x_j)(x_i - x_j)^T
\]

This paper calls \( S_C \) and \( S_M \) defined in (4) and (5) respectively as cannot-link scatter matrix and must-link scatter matrix, which resemble the concepts of between-class scatter matrix and within-class scatter matrix respectively in linear discriminant analysis (LDA). The difference lies in that the latter uses class labels to generate scatter matrices, while the former uses pairwise constraints to generate scatter matrices.
Obviously, the problem expressed by (3) is a typical Eigen-problem, and can be efficiently solved by computing the eigenvectors of $S_c - \gamma S_M$ corresponding to the largest eigenvalues. This paper considers the set of $\text{IP} = \{S_c, S_M, \gamma\}$ as Independency Parameters (IP) which are used for generating diversify basic clustering results in next stage.

3.2 Decentralization

Based on definition of WOC theory, algorithms must be able to specialize and draw on local knowledge in this step. In the proposed method, these specialize and draw on local knowledge are generated in Independency stage by reforming raw data set based on a random selected pairwise constrains, which are shown by IP. As a result, the proposed method needs a new semi-supervised basic clustering algorithms which generate basic results based on IP parameters. This paper proposes CP+K-means algorithm, which is a new branch of simple k-mean, for generating basic clustering results by using IP parameters. Fig. 2 demonstrates the pseudo code of CP+K-means algorithm:

**Algorithm** CP+K-means

**Inputs** data set $X = \{x_i\}_{i=1}^n$, Independency parameters $\text{IP} = \{\gamma, S_c, S_M\}$, Number of projective vectors $d$, Number of cluster $K$.

**Output** Partition of data set into $k$ clusters

**Method**
1. Compute the $d$ eigenvectors $W = (w_1, w_2, ..., w_d)$ of $S_c - \gamma S_M$ corresponding to the largest $d$ eigenvalues.
2. Perform classical k-means clustering on low-dimensional data set $y = \{w^T x_i\}_{i=1}^n$

*Figure 2. The CP+K-means algorithm*

As illustrated in Fig. 2, CP+K-means can reduce the size of data sets by determining the value $d$, which is number of projection vectors. In this algorithm, $X$ is raw data set; and IP parameters are generated in independency stages by (2), (4), and (5). Also, number of clusters in basic clustering results determines by users based on the subject of clustering problem, and its application.

3.3 Diversity

3.3.1 Uniformity

In Cluster Ensemble Selection methods, after generating basic results, they must be evaluated by a consensus function. In fact, tracing errors can control similarity and repetition of specific answers. Alizadeh et al. [4], [5] proposed APMM which measures the similarity between cluster $C$ and a specific partition. Unlike the NMI metric, this metric avoids the symmetry problem [4], [5], [6], [7]. Based on APMM equation, this paper proposes Uniformity as follows:
In equation (6), \( E \) is the ensemble committee (reference set), and \( Q \) is the number of partitions/members (selected basic results) in the ensemble committee; \( M \) is the number of clusters in the partition \( P \); \( n_c \) is the size of cluster \( C \) from partition \( P \); \( n \) is the number of samples which are available in the partition \( p \); and \( n_i^E \), \( K^E \) is the size of the \( i \)-th cluster of partition \( E_t \), and the number of clusters in the partition \( p_t \), respectively.

Uniformity represents the maximum value of similarity between partition \( P \) and the other partitions of ensemble committee. Since Uniformity is normalized between zero and one, we consider \( 1 - \text{Uniformity} \) to represent the diversity as follows:

\[
\text{Diversity} (p) = 1 - \text{Uniformity} (P) \geq dT \tag{7}
\]

In this paper, (7) will be called diversity condition. This condition is one of the conditions that should be established in order to appended partition to the ensemble committee is the diversity condition. In other words, if the diversity of a generated partition satisfies \( dT \) (diversity threshold), it will be added to the reference set. It is clear that for each clustering problem \( dT \) is different, and predicting optimize value for \( dT \) can be effective on performance and runtime. This paper proposes a new heuristic method for finding optimize value for diversity threshold based on supervision information, which can balance the performance and runtime of proposed method in each clustering problem. This heuristic method will be presented in the next section.

3.3.2 Threshold Value Estimation (TVE)

Thresholding is one of important parts in the CES, which can significantly change the accuracy of final result, or runtime of algorithm in different clustering problem. So, threshold value can make a balance between performance and runtime in many recent proposed methods in the CES problems. It is clear that this value always related to the type of basic clustering algorithms and the consensus function, which is used for evaluating basic clustering result. For instance, if k-means [41], [3] as a basic clustering algorithm, and NMI [3] as an evaluating metric are used for special data set, most of the time, the evaluated values are ranged between the lower and upper band numbers, which are unique for this data sets. On the one hand, if the value of threshold is determined greater than the mentioned value of upper band in that example, the set of selected partitions, which is called reference set or ensemble committee [3], [4], [5], [6], [7], [8], [10] will be empty. Furthermore, if the CES algorithm uses a greedy strategy, the runtime unusually and rapidly will be increased. On the other hand, if the threshold value is determined smaller than the value of lower band, the accuracy of final result significantly will be decreased.

For estimating best value for threshold, real class labels are needed but there is no class labels in unsupervised learning problem. Therefore, there is no clear solution in the most of unsupervised CES methods [3], [4], [5], [6], [7], [10]. Although the proposed method uses only pairwise constrain (not class labels), this paper proposes a new heuristic method, which is called Threshold Value Estimation (TVE), for
estimating the class labels of a subset of instances based on the supervision information. TVE estimates class labels of instances, which are depicted in pairwise constraint, and after that selected these instances as a new data set (a subset of raw data set); then it uses new data set and a basic clustering algorithm such as k-means for generating some basic clustering results. The maximum value of estimation between the generated basic clustering results from mention subset of original data set and estimated class labels based on pairwise constraint considers as a threshold value for diversity estimation in our proposed method. Fig. 3 illustrates the TVE pseudo code:

**Algorithm TVE**

**Inputs**
- data set \( X = \{x_i\}_{i=1}^n \)
- Pairwise constraints M and C
- Number of iterative I

**Output**
- Diversity threshold \( dT \)

**Method**

1. Generate \( \bar{X} \) as the selected instances form \( X \) based on the members of M and C.
2. Generate \( \bar{C} \) by assign same class labels for the instances of \( \bar{X} \) based on M members.
3. Fill the rest of non-labeled instances in \( \bar{C} \) based on C members by using 1-NN algorithm.
4. Generate I times basic results and store them in \( K \).
5. \( dT = \max_{k_i \in K} \{\text{ConsensusMetric}(k_i, \bar{C})\}_{i=1}^I \)  

*Figure 3. The Threshold Value Estimation (TVE)*

In Fig. 3, \( X, M, C, I, dT \) are the data set, pairwise must-link constrain, pairwise cannot-link constrain, number of iterative the basic clustering result, and the estimated threshold value, respectively. As this figure depicted, the threshold value, respectively. As this figure depicted, the instances of \( X \) selected as a new subset of data set, which is called \( \bar{X} \), if and only if there are existed on one of M or C sets. Then, \( \bar{C} \) as estimated class labels fills by assigning same class labels to the instances which has relation based on the rules of M set.

**3.4 Aggregation**

In this step, the selected basic clustering results in the ensemble committee, which is called wised crowd in this paper, are combined to reach a final consensus partition. In some of clustering method, the consensus partition uses a co-association matrix that counts the number of groupings in the same cluster for all data points. In these methods, the primary clustering results are first used to construct the co-association matrix. The most prominent of these methods is EAC [7], [23]. Each entry in the co-association matrix is computed as:

\[
C_{i,j} = \frac{n_{i,j}}{m_{i,j}}
\]  

(8)
Where \( n_{ij} \) counts the number of clusters shared by objects with indices \( i \) and \( j \); \( m_{ij} \) is the number of partitions in which this pair of objects is simultaneously present [23]. Proposed method uses the co-association matrix to aggregate the results. Then final partition is generated by using the Average-Linkage algorithm on mentioned matrix. This paper uses Average Linkage for generating Dendrogram because it has high performance in comparison with other hierarchical methods in EAC [4], [5], [7], [10]. At last, the final result is created based on clusters' number in WSCES which is described in the next section.

3.5 Summarization

Fig. 4 depicts the pseudo code of the proposed method. In this figure, \( X, NC, K, M, C, d, \) and \( I \) are the data set, number of basic clustering results which are generated for estimation, the number of clusters in the partition of final result, pairwise must-link constrains, pairwise cannot-link constrains, number of projection which is used for reducing dimension of data set, and the iterative of basic clustering algorithm. The distances are also measured by a Euclidean metric.

As illustrated in Fig. 4, \( d_T \) is calculated by TVE algorithm which is depicted in Fig. 3. Then, the half of supervision information (\( M, C \)) are selected randomly by Random-Selection function as global knowledge which are shared among all basic clustering algorithms. Next, for each basic cluttering algorithms, private knowledge randomly selected from the rest of supervision information, in equal number. After that, independency parameters (\( IP = \{\gamma, S_c, S_M\} \)) are generated by global and private knowledge; and basic clustering results are generate by using IP and CP+K-means algorithm, which is shown in Fig. 3. This procedure is repeated \( NC \)-times, which \( NC \) is the number of basic cluttering results. The EAC function builds the co-association matrix, according to (8). The Average-Linkage and Cluster functions build the final ensemble according to the Average Linkage method [4], [5], [7], [10].

The list of references is headed “References” and is not assigned a number. The list should be set in small print and placed at the end of your contribution, in front of the appendix, if one exists. Please do not insert a page break before the list of references if the page is not completely filled. An example is given at the end of this information sheet. For citations in the text please use square brackets and consecutive numbers: [1], [2], [3].
Experimental Results

This section describes a series of empirical studies and reports their results. In real world, unsupervised methods are used to find meaningful patterns in non-labeled data sets such as web documents. Since real data sets don’t have class labels, there is no direct evaluation method for evaluating the performance in unsupervised methods. Like many pervious researches [3], [4], [5], [6], [7], [10], [23] this paper compares the performance of the proposed method with other basic and ensemble methods by using standard data sets and their real classes. This paper used MATLAB R2014b (8.4) in order to generate the experimental results. As mentioned before, all distances were measured by a Euclidean metric. All results are reported as the average of 10 independent runs of the algorithm on special PC. The final clustering performance was

Algorithm WSCES

Inputs
- Dataset $X = (x_i)_{i=1}^n$
- Number of basic clustering results $NC$
- Pairwise constraints $M$ and $C$
- Number of clusters $K$
- Number of projection $d$
- Number of Iterative $I$

Output
- Predicted Class Labels index

Method

Initial Counter = 0

$dT = TVE(X, M, C, I)$

$G_c = \text{Random Selection}(C, 50\%)$

$G_M = \text{Random Selection}(M, 50\%)$

$P_c = C - G_c$

$P_M = M - G_M$

While Counter <= NC do

$R_c = \text{Random Selection}(P_c, 100/NC)$

$R_M = \text{Random Selection}(P_M, 100/NC)$

Generate $\gamma$ by using $X$, $G_c + R_c$, $G_M + R_M$ and Eq. 2. Generate $S_c$ by using $X$, $G_c + R_c$ and Eq. 4. Generate $S_M$ by using $X$, $G_M + R_M$ and Eq. 5.

Basic-Result = $CP+K$-means($G_c + R_c$, $G_M + R_M$, $\gamma$, $S_c$, $S_M$, $d$)

If (Uniformity (Basic-Result, Wised-Crowds) > $dT$) then

Add Basic-Result to Wised-Crowds

End if

Counter++

End While

Co-Association-Matrix = EAC (Wised-Crowds)

Dendrogram = Average-Linkage (Co-Association-Matrix)

Final-Result = Cluster (Dendrogram, $K$)

End function

Figure 4. The Wised Semi-Supervised Cluster Ensemble Selection (WSCES) pseudo code

4. Experimental Results

This section describes a series of empirical studies and reports their results. In real world, unsupervised methods are used to find meaningful patterns in non-labeled data sets such as web documents. Since real data sets don’t have class labels, there is no direct evaluation method for evaluating the performance in unsupervised methods. Like many pervious researches [3], [4], [5], [6], [7], [10], [23] this paper compares the performance of the proposed method with other basic and ensemble methods by using standard data sets and their real classes. This paper used MATLAB R2014b (8.4) in order to generate the experimental results. As mentioned before, all distances were measured by a Euclidean metric. All results are reported as the average of 10 independent runs of the algorithm on special PC. The final clustering performance was
evaluated by re-labeling between obtained clusters and the ground truth labels and then counting the percentage of correctly classified samples [7].

In the rest of this part, first of all, standard data sets will be introduced in details; then, the result of CP+K-means algorithm compares with some traditional unsupervised methods and classical basic semi-supervised clustering methods. After that, the performance of WSCCES compares with classical and state-of-the-art unsupervised/semi-supervised cluster ensemble methods. Next, critical parameters in the proposed method such as number of projection (d) and etc. will be analyzed. The last but not least, the effect of noisy data sets on the proposed will be studied.

4.1 Data Sets

The proposed method is applied to 3 different categories of standard data sets, which are called UCI, Handwriting and Brain Images (ADNI) data sets. We have chosen data sets which are as diverse as possible in their numbers of true classes, features, and samples, because this variety better validates the obtained results. The features of the data sets are normalized to a mean of 0 and variance of 1, i.e. N( 0,1).

4.1.1 UCI Data Sets

Table 1 shows the UCI data sets which are used in this paper. More information is available in [42].

<table>
<thead>
<tr>
<th>Name</th>
<th>Samples</th>
<th>Features</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arcene</td>
<td>900</td>
<td>10000</td>
<td>2</td>
</tr>
<tr>
<td>CNAE-9</td>
<td>1080</td>
<td>857</td>
<td>9</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>351</td>
<td>34</td>
<td>2</td>
</tr>
<tr>
<td>Optdigit</td>
<td>5620</td>
<td>62</td>
<td>10</td>
</tr>
<tr>
<td>Pendigits</td>
<td>10992</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>Sonar</td>
<td>208</td>
<td>60</td>
<td>2</td>
</tr>
<tr>
<td>Statlog</td>
<td>6435</td>
<td>36</td>
<td>7</td>
</tr>
<tr>
<td>wine</td>
<td>178</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>Yeast</td>
<td>1484</td>
<td>8</td>
<td>10</td>
</tr>
</tbody>
</table>

4.1.2 Handwriting Data Sets

4.1.2.1 USPS data

Fig. 5 depicts sample images from each class of the USPS digits data set. It is a collection of 16 x 16 grayscale images of natural handwritten digits and is available from [43], which each class contains nearly 1000 images of one of the ten digits; and each image is then represented with a 256-dimensional vector where the columns of the image are concatenated. In summarization, this data set contains 256 features, 10 classes, and 9298 samples, which is used for generating the excremental results.

Figure 5. Sample images from the USPS digits data set [36].
4.1.2.2 Persian Handwriting

This data set, which contains 1000 real handwriting, collected by Academy of Persian language and literature [44]. Each character in this language can write in 2-6 different shapes. For instance, Fig. 6 illustrates numbers from zero (left side) to nine (right side) in Persian language. As this figure shows, there are significant different in way of writing in the Persian. As a result, recognize this realworld data set can be a new alternative in modern clustering or classification methods. In addition, this data set contains 20 features, 32 classes, and 32000 samples for 16 x 16 grayscale image which is accessible in [44].

Figure 6. Sample images from numbers in the Persian Handwriting data set.

4.1.3 ADNI data set

As another real-world data set alternative, this paper uses Alzheimer's Diseases Neuroimaging Initiative (ADNI) [45] for generating the experimental results. This data set launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and nonprofit organizations, as a $60 million, 5-year public-private partnership. The main reason of collecting the ADNI data set is to understanding that serial Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of Mild Cognitive Impairment (MCI) and early Alzheimer's Diseases (AD). It is clear that categorizing patterns in this data set, which is done by unsupervised methods, or creating predictive models, which is done by supervised methods, can aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. There are two different type of images (MRI, PET) in ADNI data set, which will be described in the rest of this part.

Magnetic Resonance Imaging (MRI): In ADNI data set, raw Digital Imaging and Communications in Medicine (DICOM) MRI scans, reviewed for quality, and automatically corrected for spatial distortion caused by gradient nonlinearity and B1 field inhomogeneity. Also, these images generated by 1.5 Tesla scanners, and were collected across a variety of scanners with protocols individualized for each scanner, as defined at [45], [46].

Positron Emission Tomography (PET): In ADNI data set, PET images are acquired 30-60 min post-injection, averaged, spatially aligned, interpolated to a standard voxel size, intensity normalized, and smoothed to a common resolution of 8-mm full width at half maximum. A detailed description of PET protocols can be found at [45], [46].

This paper uses a standard sampling of ADNI data sets, which is called ADNI 202 [46]. In this sampling information of 202 subject are collected based on the clinical
experience of medical doctors. Table II shows ADNI 202 features, which is called ADNI in short from in the rest of this paper.

**Table 2: ADNI 202 Dataset [46]**

<table>
<thead>
<tr>
<th></th>
<th>AD (n=51; 18F/33M)</th>
<th>MCI (n=99; 32F/67M)</th>
<th>HC (n=52; 18F/33M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (Mean)</td>
<td>75.2</td>
<td>75.3</td>
<td>75.3</td>
</tr>
<tr>
<td>(SD)</td>
<td>7.4</td>
<td>7.0</td>
<td>7.0</td>
</tr>
<tr>
<td>Range</td>
<td>59-88</td>
<td>55-89</td>
<td>55-89</td>
</tr>
<tr>
<td>Education</td>
<td>17.7</td>
<td>15.9</td>
<td>15.8</td>
</tr>
<tr>
<td>(Mean)</td>
<td>3.6</td>
<td>2.9</td>
<td>3.2</td>
</tr>
<tr>
<td>(SD)</td>
<td>4.20</td>
<td>8.20</td>
<td>8.20</td>
</tr>
<tr>
<td>MMSE</td>
<td>23.8</td>
<td>27.1</td>
<td>29.1</td>
</tr>
<tr>
<td>(Mean)</td>
<td>2.0</td>
<td>1.7</td>
<td>1.2</td>
</tr>
<tr>
<td>(SD)</td>
<td>20-26</td>
<td>24-30</td>
<td>25-30</td>
</tr>
<tr>
<td>CDR</td>
<td>0.7</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>(Mean)</td>
<td>0.3</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>(SD)</td>
<td>0.5-1.1</td>
<td>0.5-0.5</td>
<td>0.0-0.0</td>
</tr>
</tbody>
</table>

MMSE=Mini-Mental State Examination  
CDR= Clinical Dementia Rating.

In summarization, ADNI data sets contains 202 samples, 93 features from MRI, 93 features from PET; As mention before, there are 3 classes for ADNI-T1 and 4 classes for ADNI-T2. In addition, the experimental results will be generated by the features of each type of images (MRI, PET), separately; and also combining these features together (MRI+PET) same as previous researches [46]. Fig. 7 shows a sample of MRI and PET images from ADNI.

![Figure 7. Sample images from the ADNI 202 data set [46].](image)

### 4.2 Comparison with Basic Clustering Methods

In this section, the experimental results of CP+K-means is compared with basic clustering methods and basic semi-supervised clustering methods, separately.

Firstly, the experimental result of CP+K-means is compares with K-means, FCM, Subtractive, and Single Linkage methods, as some classic unsupervised basic clustering methods. In this experiment, the number of projection (d) for CP+K-means considers 50% of number of features in each data set, and 5% of real class labels of each data set randomly selected as supervision information, which 2.5% are selected for pairwise must-link constraints and 2.5% are selected for pairwise cannot link constraint. Table 3 illustrates the performance of CP+K-means algorithm in comparison with basic clustering algorithms.
Table 3. The performance of basic clustering methods

<table>
<thead>
<tr>
<th>Data sets</th>
<th>K-means</th>
<th>FCM</th>
<th>Subtractive</th>
<th>Single Link CP+K-means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arcene</td>
<td>48.24±0.91</td>
<td>49.34±0.89</td>
<td>51.19±0.491</td>
<td>43.93±1.12</td>
</tr>
<tr>
<td>CNAE-9</td>
<td>61.9±1.53</td>
<td>62.81±0.378</td>
<td>55.89±0.295</td>
<td>18.32±0.361</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>63.5±1.342</td>
<td>64.8±0.974</td>
<td>77±0.012</td>
<td>64.38±1.304</td>
</tr>
<tr>
<td>Optdigit</td>
<td>47.23±1.241</td>
<td>38.33±0.921</td>
<td>47.72±0.312</td>
<td>10.28±0.202</td>
</tr>
<tr>
<td>Pendigits</td>
<td>40.97±1.69</td>
<td>36.77±1.02</td>
<td>10.4±0.956</td>
<td>10.46±3.92</td>
</tr>
<tr>
<td>Statlog</td>
<td>40.89±1.831</td>
<td>49.91±2.14</td>
<td>23.8±1.562</td>
<td>65.1±0.901</td>
</tr>
<tr>
<td>Wine</td>
<td>59.73±0.51</td>
<td>52.34±0.785</td>
<td>60.23±0.91</td>
<td>37.64±1.321</td>
</tr>
<tr>
<td>Yeast</td>
<td>31.19±0.692</td>
<td>29.98±0.341</td>
<td>31.2±0.57</td>
<td>31.73±0.186</td>
</tr>
<tr>
<td>USPS</td>
<td>53.29±1.326</td>
<td>54.21±0.73</td>
<td>41.91±1.69</td>
<td>23.67±3.35</td>
</tr>
<tr>
<td>Persian-HW</td>
<td>36.16±2.81</td>
<td>32.92±1.92</td>
<td>31.82±2.98</td>
<td>30.85±2.52</td>
</tr>
<tr>
<td>ADNI-MRI-T1</td>
<td>32.64±2.18</td>
<td>31.36±2.13</td>
<td>28.79±0.79</td>
<td>12.73±2.51</td>
</tr>
<tr>
<td>ADNI-MRI-T2</td>
<td>33.93±2.74</td>
<td>30.81±1.39</td>
<td>27.45±1.57</td>
<td>11.95±1.85</td>
</tr>
<tr>
<td>ADNI-PET-T1</td>
<td>31.62±2.39</td>
<td>29.61±0.82</td>
<td>26.25±0.92</td>
<td>18.31±1.52</td>
</tr>
<tr>
<td>ADNI-PET-T2</td>
<td>31.87±1.27</td>
<td>32.09±0.94</td>
<td>32.38±1.28</td>
<td>17±2.71</td>
</tr>
<tr>
<td>ADNI-FUL-T1</td>
<td>35.58±2.39</td>
<td>34.69±0.62</td>
<td>30.48±1.08</td>
<td>24.61±1.02</td>
</tr>
<tr>
<td>ADNI-FUL-T2</td>
<td>38.88±1.32</td>
<td>36.2±0.56</td>
<td>33.81±0.72</td>
<td>25.38±0.99</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>43.29±1.56</strong></td>
<td><strong>41.72±1.58</strong></td>
<td><strong>37.29±1.03</strong></td>
<td><strong>26.36±1.02</strong></td>
</tr>
</tbody>
</table>

In Table 3, the best results obtained for each dataset have been bolded. In this table, all possible types of ANDI data set is reported. The second part of ADNI data set (MRI, PET, and FUL) represents the type of images, which FUL means both features of MRI and PET are used at same time. Also T1 and T2 refers to Type 1 and Type 2, which were describe in Data Set section. As this table shows, although 50% of features were only used for generating basic results in CP+K-means, it significantly improves the performance of final result in most of the time by using 5% of real class labels as supervision information. Table 3 shows the average of performance for each technique. Even though CP+K-means was outperformed in two datasets (Yeast and Ionosphere) by some algorithms, the majority of results demonstrate superior accuracy for the proposed method. As this table depicts, the proposed method generated high performance and low errors in comparison with basic clustering algorithms.

In Fig. 8, the performance of CP+K-means compares with Constrained K-means (CK) [17] and Random Subspace (RS) [47] as classical semi-supervised clustering methods. In this experience the number of projection for CP+K-means considers 50% of number of features in each data sets. In addition, 5% to 30% of real class labels are randomly selected for generating supervision information in this experience; the half of these selected class labels were used for generating pairwise must-link constraint, and the rest of them were used for generating pairwise cannot-link constraint. As this figure shows, most of the time, the accuracy of semi-supervised methods were increased when the percentage of supervision information had been increased. Furthermore, this experiment shows that CP+K-means can generate high performance and low error basic results in comparison with classical semi-supervised clustering methods.
4.3 Comparison with Cluster Ensemble Methods

The empirical results of WSCES in comparison with some classical and state-of-the-art unsupervised and semi-supervised methods will be discussed in this section.

In Table 4, the result of WSCES compares with the EAC, APMM, MAX, and WOCCE as unsupervised cluster ensemble (selection) methods. In this experience, number of basic results in each ensemble (NC) and the number of iterative for TVE algorithm (I) for WSCES were considered 20 and 4, respectively. Also, the same numbers of basic results were assigned for other methods. Furthermore, number of projection in WSCES is 50% of number of features in each data set; and 5% of real class labels are randomly selected as supervision information (half for must-link and half for cannot-link).

Analyzing the effects of different consensus function, which are used for combing basic results, on final result is the main reason of comparing the results of WSCES with EAC (full ensemble) method. In addition, this paper compares the proposed method with APMM method because it uses an APMM based metric (same as WSCES) this method for evaluating diversity in unsupervised CES. Also, MAX method uses another metric with same name (MAX) for evaluating diversity in unsupervised CES. Analyzing the effects of different metrics, which are used for evaluating basic results, on final result is the main reason of comparing the results of WSCES with APMM and MAX methods. Furthermore, as an alternative in unsupervised CES, which uses WOC theory for generating final result, the results of WSCES are compared with WOCCE.

As Table 4 shows, although 50% of features and some of basic clustering results (after selection) were only used for generating final results of WSCES method, it improves
(1%) the performance of final result in most of the time by using 5% of real class labels as supervision information. Even though WSCES was outperformed in one dataset (Yeast) by some algorithms, the majority of results demonstrate superior accuracy for the proposed method. In the rest of this section, the empirical results of WSCES method will be compared with state-of-the-art semi-supervised methods.

Table 4. The performance of cluster ensemble methods

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Arcene</td>
<td>61.79±0.813</td>
<td>66.28±0.216</td>
<td>62.14±0.238</td>
<td>65.16±0.32</td>
<td>66.2±0.18</td>
</tr>
<tr>
<td>CNAE-9</td>
<td>74.84±0.193</td>
<td>77.42±0.792</td>
<td>78.63±0.799</td>
<td>79.2±0.579</td>
<td>80.77±0.932</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>67.8±1.118</td>
<td>70.94±0.13</td>
<td>64.48±0.914</td>
<td>70.52±0.132</td>
<td>72.34±0.98</td>
</tr>
<tr>
<td>Optdigits</td>
<td>48.12±0.503</td>
<td>77.1±0.841</td>
<td>76.11±0.65</td>
<td>77.16±0.21</td>
<td>79.17±0.1</td>
</tr>
<tr>
<td>Pendigits</td>
<td>43.9±0.43</td>
<td>47.4±0.699</td>
<td>57.02±0.521</td>
<td>58.68±0.18</td>
<td>60.9±0.56</td>
</tr>
<tr>
<td>Sonar</td>
<td>52.07±0.651</td>
<td>54.1±0.91</td>
<td>53.98±0.16</td>
<td>54.39±0.25</td>
<td>55.31±0.72</td>
</tr>
<tr>
<td>Statlog</td>
<td>43.96±0.817</td>
<td>54.88±0.528</td>
<td>54.23±0.14</td>
<td>55.77±0.719</td>
<td>60.51±0.11</td>
</tr>
<tr>
<td>Wine</td>
<td>70.56±0.89</td>
<td>64.6±0.231</td>
<td>69.17±0.789</td>
<td>71.34±0.542</td>
<td>72.32±0.034</td>
</tr>
<tr>
<td>Yeast</td>
<td>31.74±0.234</td>
<td>31.06±0.245</td>
<td>32.4±0.124</td>
<td>32.76±0.268</td>
<td>32.93±0.109</td>
</tr>
<tr>
<td>USPS</td>
<td>62.8±0.69</td>
<td>63.91±0.94</td>
<td>64.73±0.48</td>
<td>65.21±0.69</td>
<td>68.26±0.521</td>
</tr>
<tr>
<td>Persian-HW</td>
<td>44.18±0.72</td>
<td>45.92±0.61</td>
<td>46.22±0.92</td>
<td>47.43±0.73</td>
<td>49.45±0.61</td>
</tr>
<tr>
<td>ADNI-MRI-T1</td>
<td>42.19±0.37</td>
<td>48.01±0.56</td>
<td>49.62±0.17</td>
<td>48.82±0.37</td>
<td>51.17±0.98</td>
</tr>
<tr>
<td>ADNI-MRI-T2</td>
<td>39.5±0.31</td>
<td>39.93±0.29</td>
<td>41.81±0.45</td>
<td>40.22±0.44</td>
<td>42.11±0.78</td>
</tr>
<tr>
<td>ADNI-PET-T1</td>
<td>40.48±0.52</td>
<td>48.37±0.82</td>
<td>47.92±0.37</td>
<td>49.19±0.26</td>
<td>50.28±0.52</td>
</tr>
<tr>
<td>ADNI-PET-T2</td>
<td>38.85±0.59</td>
<td>38.53±0.17</td>
<td>37.83±0.29</td>
<td>39.43±0.79</td>
<td>41.59±0.16</td>
</tr>
<tr>
<td>ADNI-FUL-T1</td>
<td>44.42±0.91</td>
<td>47.22±0.93</td>
<td>46.62±0.38</td>
<td>48.82±0.41</td>
<td>49.52±0.35</td>
</tr>
<tr>
<td>ADNI-FUL-T2</td>
<td>47.21±0.63</td>
<td>50.09±0.35</td>
<td>48.54±0.14</td>
<td>49.39±0.63</td>
<td>51.21±0.32</td>
</tr>
<tr>
<td>Average</td>
<td>50.26±0.61</td>
<td>54.46±0.54</td>
<td>54.79±0.44</td>
<td>56.09±0.44</td>
<td>57.3±0.47</td>
</tr>
</tbody>
</table>

Pairwise Constraints=5%, Number of basic clustering results NC=20, Number of iterative in TVE I=4.

In Table 5, the result of WSCES compares with the Random Projection (RP) [47] as classical semi-supervised cluster ensemble method and BGCM [46], SKMS, and NBF as state-of-the-art semi-supervised cluster ensemble (selection) methods. In this experience, number of basic results in each ensemble (NC) and the number of iterative for TVE algorithm (I) for WSCES were considered 20 and 4, respectively. Also, the same numbers of basic results were assigned for other methods. Furthermore, number of projection in WSCES is 50% of number of features in each data set; and 30% of real class labels are randomly selected as supervision information (half for must-link and half for cannot-link). These supervision information are unique in each run (there are 10-times run for each empirical results) for all mentioned methods.

In Table 5 the best results obtained for each dataset have been bolded. As this table shows, although WSCES was outperformed in three datasets (Wine, Yeast, and ADNI-MRI-T1) by some algorithms, the majority of results demonstrate superior accuracy for the proposed method. Moreover, it is difficult to separate the WSCES and NBF methods. However, the average performance over all seventeen datasets reveals that WSCES outperformed NBF by over 2%. Needless to say that both methods use same background especially in diversity evaluation, and unsupervised clustering concepts which are used in this method [3, 4, 5, 6, 7, 30].

In Fig. 9, the performance of WSCES compares with RP, BGCM, SKMS, and NBF. In this experience the number of projection for WSCES considers 50% of number of features in each data sets. In addition, 5% to 30% of real class labels are randomly selected for generating supervision information in this experience; the half of these selected class labels were used for generating pairwise must-link constraint, and the rest of
them were used for generating pairwise cannot-link constraint. As this figure shows, most of the time, the accuracy of semi-supervised methods were increased when the percentage of supervision information had been increased. Furthermore, this experiment shows that WSCES can generate high performance and low error basic results in comparison with other methods.

### Table 5. The performance of Semi-supervised cluster Ensemble methods

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Arcene</td>
<td>60.3±0.3</td>
<td>66.3±0.18</td>
<td>69.1±0.498</td>
<td>67.1±0.93</td>
<td>70.2±0.18</td>
</tr>
<tr>
<td>CNAE-9</td>
<td>69.2±0.361</td>
<td>80.1±0.459</td>
<td>82.7±0.12</td>
<td>79.9±0.812</td>
<td>85.7±0.932</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>68.4±0.624</td>
<td>73.6±0.341</td>
<td>72.6±0.342</td>
<td>70.8±0.652</td>
<td>74.3±0.98</td>
</tr>
<tr>
<td>Optdigits</td>
<td>67.2±0.554</td>
<td>71.5±0.692</td>
<td>76.9±0.274</td>
<td>78.4±0.21</td>
<td>82.1±0.1</td>
</tr>
<tr>
<td>Pendigits</td>
<td>59.3±0.481</td>
<td>65.1±0.42</td>
<td>67.7±0.591</td>
<td>64.2±0.614</td>
<td>68.9±0.56</td>
</tr>
<tr>
<td>Sonar</td>
<td>52.6±0.313</td>
<td>52.0±0.873</td>
<td>56.2±0.872</td>
<td>55.2±0.413</td>
<td>56.3±0.72</td>
</tr>
<tr>
<td>Statlog</td>
<td>64.5±0.737</td>
<td>65.7±0.591</td>
<td>63.3±0.42</td>
<td>70.1±0.144</td>
<td>72.5±0.11</td>
</tr>
<tr>
<td>Wine</td>
<td>69.8±0.427</td>
<td>72.4±0.141</td>
<td>73.9±0.762</td>
<td>71.1±0.621</td>
<td>72.3±0.034</td>
</tr>
<tr>
<td>Yeast</td>
<td>19.7±0.311</td>
<td>28.1±0.462</td>
<td>25.6±0.529</td>
<td>30.8±0.226</td>
<td>24.9±0.109</td>
</tr>
<tr>
<td>USPS</td>
<td>61.3±0.21</td>
<td>68.3±0.24</td>
<td>69.0±0.61</td>
<td>66.7±0.21</td>
<td>72.2±0.521</td>
</tr>
<tr>
<td>Persian-HW</td>
<td>41.8±0.39</td>
<td>48.3±1.91</td>
<td>49.8±0.69</td>
<td>53.2±0.79</td>
<td>55.4±0.61</td>
</tr>
<tr>
<td>ADNI-MRI-T1</td>
<td>41.5±0.23</td>
<td>55.4±0.83</td>
<td>49.3±0.31</td>
<td>53.7±0.37</td>
<td>55.1±0.98</td>
</tr>
<tr>
<td>ADNI-MRI-T2</td>
<td>35.7±0.51</td>
<td>44.3±0.51</td>
<td>42.6±0.44</td>
<td>45.0±0.51</td>
<td>46.1±0.78</td>
</tr>
<tr>
<td>ADNI-PET-T1</td>
<td>42.3±0.38</td>
<td>52.0±0.71</td>
<td>50.9±0.26</td>
<td>52.6±0.47</td>
<td>53.2±0.52</td>
</tr>
<tr>
<td>ADNI-PET-T2</td>
<td>37.8±0.64</td>
<td>43.6±0.93</td>
<td>42.3±0.79</td>
<td>43.7±0.52</td>
<td>45.5±0.16</td>
</tr>
<tr>
<td>ADNI-FUL-T1</td>
<td>40.3±0.91</td>
<td>55.2±1.03</td>
<td>54.8±0.41</td>
<td>56.4±0.71</td>
<td>57.5±0.35</td>
</tr>
<tr>
<td>ADNI-FUL-T2</td>
<td>40.2±0.23</td>
<td>54.4±0.83</td>
<td>52.4±0.63</td>
<td>55.1±0.57</td>
<td>56.2±0.32</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>51.4±0.7</strong></td>
<td><strong>55.6±0.66</strong></td>
<td><strong>58.8±0.5</strong></td>
<td><strong>59.7±0.52</strong></td>
<td><strong>61.7±0.47</strong></td>
</tr>
</tbody>
</table>

Pairwise Constraints=30%, Number of basic clustering results NC=20, Number of iterative in TVE I=4.

---

**Figure 9. The performance of semi-supervised cluster ensemble methods**

As summarization, based on this section experience it can be say that WSCES can generate more stable results in comparison with other methods because it uses a robust
framework (WOC theory) for controlling the process of generating, evaluating, selecting and combing clustering results same as two other methods in (un)supervised learning [7, 24].

5. Discussion

Semi-supervised clustering which use supervision information usually in the form of pairwise constraints have been well-developed in recent years. However, one important yet remaining unresolved problem in semi-supervised clustering is how to choose the most appropriate constraints set for specific algorithms and tasks. This paper suggests a new method for employing the Wisdom of Crowds, which is a theory in social science, for proposing a robust framework in Semi-Supervised Cluster Ensemble Selection. The most important advantage of this employment is the addition of new aspects, such as independency and decentralization, which are based on the concepts of features selection and eigenvectors, as well as a new strategy for selecting high quality basic clustering results. Also this paper proposes a new algorithm to predict the diversity threshold value, which is called TVE method, in the process of selecting basic results. We also introduce the Uniformity criterion to measure the diversity of the basic results.

To prove the claims of this paper, the results of the proposed method are compared with the results of (semi-supervised) clustering methods, cluster ensemble (selection) methods, and state-of-the-art semi-supervised cluster ensemble (selection) methods. The results were achieved by applying the mentioned methods on different kinds of standard data sets, which categorized in UCI, Handwriting, and Brain Image (ADNI) groups. Furthermore, this paper introduces a new real-world data set for recognizing Persian (Iranian Language) handwriting. In our empirical results, data sets with different scales (small, average, and large) were used so that the accuracy could be evaluated regardless of the scale of a data set. In addition, in order to be ensured about the accuracy of all results, the experiment has been repeated 10 times. Similar to other pioneering ideas, the proposed framework can be improved later. This paper suggests different semi-supervised clustering algorithms in the structure of proposed framework. In addition, analyzing parameters and basic components in the proposed algorithms for independency, decentralization, diversity and threshold estimation (TVE) can be another challenge for the feature works.

References


