2015 IEEE International Conference on Data Mining

Weighted Spectral Cluster Ensemble

Muhammad Yousefnezhad
Department of Computer Science and Technology
Nanjing University of Aeronautics and Astronautics
Nanjing, China
myousefnezhad@nuaa.edu.cn

Daoqiang Zhang
Department of Computer Science and Technology
Nanjing University of Aeronautics and Astronautics
Nanjing, China
dqzhang@nuaa.edu.cn

***Abstract*—Clustering explores meaningful patterns in the non-labeled data sets. Cluster Ensemble Selection (CES) is a new approach, which can combine individual clustering results for increasing the performance of the final results. Although CES can achieve better final results in comparison with individual clustering algorithms and cluster ensemble methods, its performance can be dramatically affected by its consensus diversity metric and thresholding procedure. There are two problems in CES: 1) most of the diversity metrics is based on heuristic Shannon's entropy and 2) estimating threshold values are really hard in practice. The main goal of this paper is proposing a robust approach for solving the above mentioned problems. Accordingly, this paper develops a novel framework for clustering problems, which is called Weighted Spectral Cluster Ensemble (WSCE), by exploiting some concepts from community detection arena and graph based clustering. Under this framework, a new version of spectral clustering, which is called Two Kernels Spectral Clustering, is used for generating graphs based individual clustering results. Further, by using modularity, which is a famous metric in the community detection, on the transformed graph representation of individual clustering results, our approach provides an effective diversity estimation for individual clustering results. Moreover, this paper introduces a new approach for combining the evaluated individual clustering results without the procedure of thresholding. Experimental study on varied data sets demonstrates that the prosed approach achieves superior performance to state-of-the-art methods.**

***Keywords*-cluster ensemble; spectral clustering; normalized modularity; weighted evidence accumulation clustering**

I. Introduction

Clustering, the art of discovering meaningful patterns in the non-labeled data sets, is one of the main tasks in machine learning. Generally, individual clustering algorithms provide different accuracies in a complex data set because they generate the clustering results by optimizing a local or global function instead of natural relations between data points in each data set. [1], [2]. As a novel solution, cluster ensemble which combines the different clustering results was proposed for achieving a better final result [1]. Cluster Ensemble Selection (CES) is a new solution which combines a selected group of best individual clustering results according to consensus metric(s) from ensemble committee in order to improve the accuracy of final results [3]. The evaluation metric(s), thresholding and selection strategy, and aggregation method are the most important challenges in CES for selecting better partitions of ensemble committee and generating the final result. There are a wide range of ideas for solving mentioned challenges [3], [4], [5], [6], [7]. Although these methods can improve performance and robustness of final results, using a wide range of threshold values and employing the entropy based metric are two main weak points of this method. Threshold values are different for each data set in the mentioned methods; and it is really hard to find optimum values in real-world applications. Moreover, most of the real-world data sets do not have logarithm behavior. So, there is no prove that entropy based methods, which estimate the diversity based on the logarithm, were the best choice to evaluate the diversity. This paper proposes a novel methodology for solving clustering problems *without mentioned weak points*.

As mentioned before, there are four stages in Cluster Ensemble Selection (CES); i.e. generating individual clustering results, evaluating, selecting and combining them as a final clustering result. Although CES can achieve a better result in comparison with individual clustering algorithms and cluster ensemble methods, the accuracy of CES is fully sensitive to the process of thresholding for selecting individual clustering results, and the consensus metric, which is used for diversity or quality estimation of the results. Unfortunately, it is so hard to find the optimum threshold values in practice; and most of the metrics, which were used for diversity or quality estimation, are heuristic; especially they are based on Shannons entropy. The main goal of this paper is solving mentioned problems. This paper proposes a new method for estimating the diversity of generated individual clustering results by using a redefined version of modularity, which is based on expected value and it is introduced for the community detections applications. Further, this paper introduces a novel approach for combining the evaluated individual clustering results without the process of thresholding.

Our contribution in this paper can be summarized as follows: *Firstly,* this study proposes a greedy method based on feedback mechanism [8] which employs the idea of bisecting k-means for generating individual results. *After that,* this paper introduces the Two Kernels Spectral Clustering (TKSC) for generating individual clustering results.

1550-4786/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDM.2015.145

This algorithm generates hybrid individual clustering results, which contains Partitional results and Modular results. Same as simple clustering problems, our method generates Partitional results; and also it generates Modular results, which represented by a graph, as a new alternative for evaluating and combining the individual results. *Next,* to satisfy the diversity criterion, this study proposes Normalized Modularity, which is a redefined version of Modularity criterion in community detection [9], for evaluating diversity of individual results in the general clustering problems. Unlike most of the diversity metrics which are based on Shannon's entropy, this metric uses Expected Value in probabilistic theory for evaluating individual clustering results and avoids the undesired logarithm [9], [10]. *Lastly,* this paper proposed Weighted Evidence Accumulation Clustering (WEAC) to obtain the final clustering with a weighted combination of all individual results. While the weight of each individual result in WEAC can be estimated with different metrics, the normalized modularity was used in this paper.

The rest of this paper is organized as follows: In Section 2, this study first briefly reviews some related works on cluster ensemble selection. Then, it introduces the proposed Weighted Spectral Clustering Ensemble (WSCE) framework in Section 3. Experimental results are reported in Section 4; and finally this paper presents conclusion and pointed out some future works in Section 5.

II. Related Works

As an unsupervised method, Clustering discovers meaningful patterns in the non-labeled data sets. There is a wide range of studies, which try to increase the performance of clustering algorithms. For instance, Zhang et al. introduced a multi-manifold regularized nonnegative matrix factorization framework (MMNMF) which can preserve the locally geometrical structure of the manifolds for multi-view clustering [11]. Anyway, individual clustering algorithms provide different accuracies in a complex data set because they generate the clustering results by optimizing a local or global function instead of natural relations between data points in each data set [1], [2].

Generally, a cluster ensemble has two important steps: Firstly, generating individual clustering results by using different algorithms and changing the number of their partitions. Then, combining the primary results and generating the final ensemble. This step is performed by consensus functions (aggregating mechanism) [1], [12].

The idea that not all partitions are suitable for cooperating to generate the final clustering was proposed in CES [3]. Instead of combing all achieved individual results, CES can combine a selected group of best individual results according to consensus metric(s) from the ensemble committee in order to improve the accuracy of final results [3], [5], [8], [4], [7]. Fern and Lin developed a method to effectively select individual clustering results for ensemble and the final decision [3]. Azimi et al. proved that diversity maximization is not an effective approach in some real-world applications. They explored that the thresholding procedure must be done based on the complexity and quality of data sets [4]. Jia et al. proposed SIM for diversity measurement, which works based on the Normalized Mutual Information (NMI) [6]. Romano et al. proposed Standardized Mutual Information (SMI) for evaluating clustering results [13].

Yousefnezhad et al. introduced independency metric instead of quality metric for evaluating the process of solving a problem in the CES [7]. Alizadeh et al. have concluded the disadvantages of NMI as a symmetric criterion. They used the APMM[1] and Maximum (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 [5], [8]. Alizadeh et al. introduced Wisdom of Crowds Cluster Ensemble (WOCCE), which is a novel method base on a theory in social science [8]. Although, this method can generate high performance and more stable results in comparison with other CES methods, using a wide range of thresholds and employing different types of clustering algorithms for generating individual results are two main problems in this method. Alizadeh et al. used A3, which is based on Shannon's entropy, for diversity evaluation; and Basic Parameter Independency (BPI), which uses initialized values of individual clustering algorithms such as random seeds in the first iterative of k-means, for independency evaluation. In addition, they introduced the feedback mechanism for generating the high-quality results [8].

As a graph based clustering methods, spectral clustering generates high-performance results when it is applied to different applications; i.e. from image segmentation to community detection arena. Kuo et al. introduced a new method for automating the process of Laplacian creation in the medical applications; especially for fMRI segmentation where this method used standard Laplacians perform poorly [14]. Chen et al. proposed a clustering algorithm which is based graph clustering and optimizing an appropriate weighted objective, where larger weights are given to observations (edge or no-edge between a pair of nodes) with lower uncertainty [15]. Gao et al. introduced a graph-based consensus maximization (BGCM) method for combining multiple supervised and unsupervised models. This method consolidated a classification solution by maximizing the consensus among both supervised predictions and unsupervised constraints [16].

III. The proposed method

Given a set of high-dimensional data examples $\hat{X} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n\}$. The simple average of $\hat{X}$ can be denoted

[1] Alizadeh-Parvin-Moshki-Minaei

as follows:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i \tag{1}$$

where n is the number of instances in the $\hat{X}$; and $\hat{x}_i$ denotes the $i-th$ instance of the data points. At the beginning, this paper minimized the correlation between features. So, it denotes X as follows:

$$X = \hat{X} - \bar{X} = \{(\hat{x}_1 - \bar{x}_1), (\hat{x}_2 - \bar{x}_2), \ldots, (\hat{x}_n - \bar{x}_n)\} \tag{2}$$

where $\hat{X}$ is the data points, and $\bar{X}$ denotes simple average of $\hat{X}$, which calculated by (1). It's clear that X is zero-mean. In other words, the excepted value of X is zero as follows:

$$\mathbb{E}\{X\} = 0 \tag{3}$$

Now, this paper maps $Q : X \in \mathbb{R}^{m\times n} \rightarrow Y \in \mathbb{R}^{m\times n}$, where m, n denote the number of features and data points, respectively. This mapping just minimizes the correlation between features. This problem can be reformulate as follows:

$$Y = Q^T X \tag{4}$$

If the correlation (covariance) of X is considered $R = \mathbb{E}\{XX^T\} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$, then the correlation of Y will be defined as follows:

$$\begin{aligned}\mathbb{E}\{YY^T\} = \mathbb{E}\{(Q^T X)(Q^T X)^T\} = \\ \mathbb{E}\{Q^T XX^T Q\} = Q^T\mathbb{E}\{XX^T\}Q = Q^T RQ\end{aligned} \tag{5}$$

Based on above definition, the expected value of $j-th$ feature of X denotes as follows:

$$\mathbb{E}\{Y_j Y_j^T\} = q^T Rq \tag{6}$$

where q denotes the $j-th$ index of the Q. In other words, our correlation problem is changed to a variance probe. Now, maximizing the q based on the variance of X will be omitted the correlation between features. Since the scale of data after mapping must be same, we assume following equation:

$$\|q\| = 1 \tag{7}$$

For maximizing the (6), which is denoted by $\Psi(q)$, our problem will be reformulated as follows:

$$\begin{aligned}max[\Psi(q) = q^T Rq] \Rightarrow \\ \frac{\partial\Psi(q)}{\partial q} = 0 \Rightarrow \\ \Psi(q+\delta q) = \Psi(q) \Rightarrow \\ (q+\delta q)^T R(q+\delta q) = q^T Rq\end{aligned} \tag{8}$$

where the symbol δq is an abbreviation for 'a small change in q'. We consider $(\delta q)^T\delta q \approx 0$, so the above definition denotes as follows:

$$(\delta q)^T Rq = 0 \tag{9}$$

Based on (7) and (8), we can assume as follows:

$$\|\delta q - q\| = \|q\| = 1 \Rightarrow (\delta q)^T q = 0 \tag{10}$$

Now, this paper defines following equation by using (9) and (10):

$$\begin{aligned}(\delta q)^T Rq - \lambda(\delta q)^T q = 0 \Rightarrow \\ (\delta q)^T[Rq - \lambda q] = 0\end{aligned} \tag{11}$$

where $\lambda \in \mathbb{R}$ is a constant. Since $(\delta q)^T \neq 0$, the following equation must be satisfy for minimizing correlation between features:

$$Rq = q\lambda \tag{12}$$

where R and λ denotes the eigenvectors and eigenvalues, respectively. For all features of X the above equation will be denoted as follows:

$$RQ = Q\Lambda \tag{13}$$

which is called eigenstructure equation. In above equation, Λ is a diagonal matrix. Based on (7), we can define following equation:

$$\|q\|^2 = 1 \Rightarrow Q^T Q = \mathbb{I} \tag{14}$$

where $\mathbb{I}$ is identity matrix. Following equation denotes based on (13) and (14):

$$\begin{aligned}RQ = Q\Lambda \Rightarrow \\ RQQ^T = Q\Lambda Q^T \Rightarrow \\ R\mathbb{I} = Q\Lambda Q^T \Rightarrow \\ R = Q^T\Lambda Q \Rightarrow \\ R = \sum_{j=1}^{m}\lambda_j q_i q_j^T\end{aligned} \tag{15}$$

where m denotes number of features in data X. Now, consider that R is a descending order based on Λ values. For an optional feature selection we can define the following equation instead of (15):

$$R = \sum_{j=1}^{d}\lambda_j q_i q_j^T \tag{16}$$

where $d < m$ is the number of features, which must be selected for generating results. Algorithm 1 shows the mapping function, which can minimized correlation of data set based on above definitions.

For generating individual clustering results, the proposed method partitions Y into C^l clusters, where k denotes number of clusters in the individual results, and C^l is $l-th$ individual result in the reference set. This paper uses the range of $l \in [2, k+2]$ for generating individual results, where k is the number of clusters in the final result. This is the same as bisect k-means algorithms but instead of applying the algorithm on generated results in each iterative, our proposed method stores this result on the ensemble committee; and then evaluates and combines these results. In other words, the reference set denotes $\zeta \in \mathbb{R}^{n\times[2,k+1]} = \{C^l\} = \{C^2, \ldots, C^{k+2}\}$.

Algorithm 1 The Mapping Function

Input: Data set $\hat{X} \in \mathbb{R}^{m \times n} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n\}$,
d as number of features:
$d = 0$ *is considered for deactivating the feature selection*
Output: Mapped data set Y
Method:
1. Calculating simple average $\bar{X}$ by using (1).
2. Calculating X by using (2).
3. Calculating $R = \mathbb{E}\{XX^T\} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$.
4. Calculating Λ and Q as eigenvalues/vectors of R.
5. Sorting Q based on descending values of λ.
6. **if** d is not zero ($d \neq 0$)
then selecting $[1, d]$ features of Q, and sorting as Q_d,
else $Q_d = Q$, $d = m$.
end if
7. Return $Y = Q_d^T X$.

Like other spectral methods, this paper calculates the non-symmetric distances (adjacency) matrix of Y, which is denoted by A [17], [18]. In the rest of this paper, our proposed method will be applied to the matrix A for each individual clustering results. Moreover, this paper uses (17) as transform function for converting distances matrix A to similarity matrix S. This transformation can optimize the memory usage [17], [18].

$$S_{i,j} = \begin{cases} exp\left(\frac{-\|y_i - y_j\|_2}{\phi^2}\right) & if\ i \neq j \\ 0 & if\ i = j \end{cases} \quad (17)$$

where y_n denotes the $n-th$ data point and $\|y_i - y_j\|_2$ will be calculated by Euclidean distance. The scaling parameter ϕ controls how rapidly affinity $S_{i,j}$ falls off with the distance between the data points. This paper uses Ng et al. method for estimating this value automatically (count non-zero values in each columns of distance matrix A) [17], [18].

This paper introduces Two Kernels Spectral Clustering (TKSC) algorithm, which can generate all individual results (ζ). Unlike normal clustering algorithms, which just generate a partition as the result, the TKSC algorithm generates two independent consequences, which are called Partitional result and Modular result, for each of the individual clustering results by using two kernels ($C^l = \{P^l, M\}$). Partitional result (P^l) is a partitioning of data points same as the result of other clustering methods; and Modular result (M) is a network of data points, which can be represented by a graph. This paper uses Modular result as a reference for evaluating the diversity of generated partition by using community detection methods [9], [10]. Furthermore, kernel in the TKSC refers to Laplacian equation in spectral methods because it transforms data points in new environment, especially linear environment for non-linear data sets.

Partitional Kernel: This paper uses following equation for generating Partitional result:

$$L_P = \mathbb{I} - D^{1/2} S D^{1/2} \quad (18)$$

where $\mathbb{I}$ is the identity matrix [17]; D is the diagonal matrix of S ($D = diag(S)$); and S will be calculated by (17). As shows in follows, the eigendecomposition is performed for calculating eigenvectors of L_P:

$$V = eigens(L_P) \quad (19)$$

where the matrix V is the eigenvectors of Partitional Kernel. The coefficient W will be defined for normalizing the matrix V:

$$W_i = \left(\sum_{i=1}^{n} V_{i1} \times V_{i2}\right)^{\frac{1}{2}} + \epsilon \quad (20)$$

where V_{ij} shows the i-th row and j-th column of the matrix V; and ϵ is used for omitting the effect of zeros in the matrix W. This paper uses $\epsilon = 10^{-20}$ for generating the experimental results. Also, n denotes the number of instances in the data set ($W \in \mathbb{R}^n$). The normalized matrix of eigenvectors will be calculated as follows:

$$U_{ij} = V_{ij} \times W_i \quad (21)$$

where U_{ij} and V_{ij} denote the i-th row and j-th column of these matrices; and W_i is the i-th row of the matrix W which is used for normalization. The Partitional result of TKSC will be calculated by applying the simple k-means [8] on the matrix U as follows:

$$P^l = kmeans(U, l) \quad (22)$$

where K is the number of classes in individual results; and U will be calculated by (21).

Modular Kernel: This paper uses following equation for generating Modular result:

$$L_M = D - S \quad (23)$$

where D is the diagonal matrix of S ($D = diag(S)$); and S will be calculated by (17). This paper considers the normalized matrix of L_M an adjacency matrix of graph representation of individual result as follows:

$$M = \frac{1}{max(L_M)} L_M \quad (24)$$

where L_M is calculated by (23), and the function max finds the biggest value in the matrix L_M. Further, all values in the matrix M, which is called Modular result, are between zero and one. Algorithm 2 shows the pseudo code of the TKSC method. Tracing errors can control similarity and repetition of specific answers in clustering problems. There is a wide range of metrics, which are based on Shannon's entropy[8], [5], for evaluating the diversity of individual results in the CES methods, such as MI [1], NMI [12], APMM [5], MAX [8], and SMI [13]. Shannon's entropy uses the logarithm of probability of individual results for

Algorithm 2 Two Kernels Spectral Clustering (TKSC)

Input: Distance matrix A, Number of clusters l
Output: Partitional result P^l, Modular result M
Method:

1. Generate similarity matrix S by using A on (17).
2. Generate diagonal matrix D by using S.
3. Generate L_P by applying S and D on (18).
4. Generate L_M by using S and D on (23).
5. Generate the matrix V as eigenvectors of L_p.
6. Generate U as normalized V by using (20) and (21).
7. Generate M by applying L_M on (24).
8. $P^l = kmeans(U, l)$
9. Return P^l and M

evaluating the diversity but there is no mathematical prove that all real-world data sets have logarithmic behavior. In community detection arena [9], [10], Modularity, which is based on Expected Value, was proposed for solving this problem. Recently, many papers proved that modularity [9], [10] can estimate the diversity on graph data sets better than entropy based methods. Unfortunately, modularity can measure the diversity only for graph data [9]. This paper proposes TKSC, which can generate a graph based result, called Modular result, for any types of data sets in real-world application. Since modularity was defined for community detection arena, this paper introduces a redefined version of modularity metric for general clustering problems, which is called Normalized Modularity (NM). It is used for evaluating the diversity of the individual results based on Modular result of the TKSC as follows:

$$NM(P^l, M) = \frac{1}{2} + \frac{1}{4z} \sum_{ij} \left[\Gamma_{ij} - \frac{\sigma_i \sigma_j}{2z} \right] \Theta(c_i, c_j) \quad (25)$$

where P^l and M are calculated by (22) and (24), respectively; z is sum of all cells in the matrix M ($z = \sum_M M_{ij}$); and c_i and c_j are the cluster's numbers of the i-th and j-th instances in the Partitional result P^l. Also, σ_i and σ_j show the degree of i-th and j-th nodes in the graph of matrix M (How many rows contains non-zero value in the columns i or j). In addition Γ_{ij} and $\Theta(c_i, c_j)$ will be calculated as follows:

$$\Gamma_{ij} = \begin{cases} 0 & \textit{if } M_{ij} = 0 \\ 1 & \textit{Otherwise} \end{cases} \quad (26)$$

$$\Theta(c_i, c_j) = \begin{cases} 1 & \textit{if } c_i = c_j \\ 0 & \textit{Otherwise} \end{cases} \quad (27)$$

This diversity evaluation is $0 \leq NM \leq 1$. In the rest of this section, we describe how NM will be used for evaluating individual clustering results. Thresholding is used for selecting the evaluated individual results in the CES. Then *co-association* matrix is generated by using consensus function on the selected results. Lastly, the final result is generated by applying linkage methods on the co-association

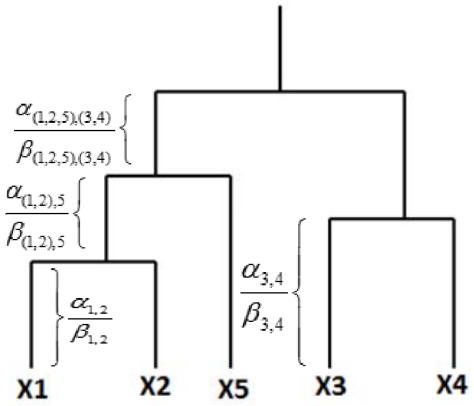


Figure 1. In the traditional EAC, the $\alpha_{(i,j)}$ represents the number of clusters shared by objects with indices (i, j); and $\beta_{(i,j)}$ is the number of partitions in which this pair of instances (i and j) is simultaneously presented. This method assumes the weights of all individual clustering results ($\alpha_{(i,j)}$) are the same. This paper proposes Weighted EAC for optimizing this method by using a weight for each individual clustering results instead of just counting their shared clusters. While the weight can have different definitions in the other applications, this paper uses average of Normalized Modularity (NM) of two algorithms as the weight in the WEAC ($\bar{\alpha}_{(i,j)} = \sum_{\alpha(i,j)} \rho_{i,j}$).

matrix. These methods generate the Dendrogram and cut it based on the number of clusters in the result [12], [8]. In recent years, many papers have used EAC as a high-performance consensus function for combining individual results [12], [5], [8], [4], [3]. EAC uses the number of clusters shared by objects over the number of partitions in which each selected pair of objects is simultaneously presented for generating each cell of the co-association matrix. Figure 1 illustrates the effect of the EAC equation ($c(i,j) = \frac{\alpha(i,j)}{\beta(i,j)}$) on the shape of Dendrogram. Where $\alpha_{(i,j)}$ represents the number of clusters shared by objects with indices (i, j); and $\beta_{(i,j)}$ is the number of partitions in which this pair of instances (i and j) is simultaneously presented. As a matter of fact; EAC considers that the weights of all algorithms results are the same. Instead of counting these indices, this paper uses following equation, which is called Weighted EAC (WEAC), for generating the co-association matrix.

$$c(i,j) = \frac{\sum_{\alpha(i,j)} \rho_{i,j}}{\beta(i,j)} \quad (28)$$

where $\alpha(i,j)$ and $\beta(i,j)$ are same as the EAC equation; Also, $\rho_{i,j}$ is the weight of combining the instances. Although this weight can have different definitions in the other applications, this paper uses average of Normalized Modularity of two algorithms as follows for combining individual results:

$$\rho_{ij} = \frac{1}{2}(NM_i + NM_j) \quad (29)$$

where NM_i and NM_j illustrates the Normalized Modularity of the algorithms, which generate the results for indices i and j. In other words, as a new mechanism, this paper generates the effective results when both algorithms

have high NM values; and also the effects of individual results are near of zero when the both algorithms have small values in the NM metric. As a result, this paper just omits the effect of low quality individual results by using mentioned mechanism instead of selecting them by thresholding procedures. Further, the final co-association matrix, which is a symmetric matrix, will be generated by (28) as follows:

$$\xi = WEAC(\zeta) = \begin{pmatrix} c(1,1) & c(1,2) & \dots & c(1,n) \\ c(2,1) & c(2,2) & \dots & c(2,n) \\ \vdots & \vdots & \vdots & \vdots \\ c(i,1) & c(i,2) & c(i,j) & c(i,n) \\ \vdots & \vdots & \vdots & \vdots \\ c(n,1) & c(n,2) & \dots & c(n,n) \end{pmatrix} \tag{30}$$

where n is the number of data points; and $c(i,j)$ denotes the final aggregation for $i-th$ and $j-th$ instances. Algorithm 3 illustrates the pseudo code of the proposed method. In this algorithm, $\hat{X}$ is the data set; k is the number of clusters in the final result; P_f is the final result partition. The distances are also measured by an Euclidean metric. The TKSC function builds the partitions and modules of individual results; and NM function evaluates these results by using (25). Then, the evaluated results will be added to reference set. The WEAC function generates the co-association matrix, according to (28), by using the Normalized Modularity values and Partitional results. The Average-Linkage function creates the final ensemble according to the average linkage method [5], [8].

Algorithm 3 The Weighted Spectral Cluster Ensemble

Input:
Data points $\bar{X}$,
Number of clusters k,
Number of features d.
Output: final result P_f
Method:
1. Generate Y by using $\bar{X}$ and d on Algorithm 1.
2. Generate matrix A by using Y based on [17].
3. **for** $l = 2$ **to** $k+2$ **do**
4. $\left[P^l, M\right] = TKSC\,(A, l)$ by using Algorithm 2.
5. $Q = NM(P^l, M)$ based on (25)
6. Add $[P, Q]$ to ζ as the reference set.
7. **end for**
8. Generate co-association matrix $\xi = WEAC(\zeta)$
9. $P_f = Average - Linkage(\xi)$

IV. Experiments

The empirical studies will be presented in this section. The unsupervised methods are used to find meaningful patterns in non-labeled datasets such as web documents, etc. in real world application. Since the real dataset doesnt have class labels, there is no direct evaluation method for estimating the performance in unsupervised methods. Like many pervious researches [12], [3], [5], [8], [7], this paper compares the performance of its proposed method with other individual clustering methods and cluster ensemble (selection) methods by using standard datasets and their real classes. Although this evaluation cannot guarantee that the proposed method generate high performance for all datasets in comparison with other methods, it can be considered as an example for analyzing the probability of predicting good results in the WSCE.

The results of the proposed method are compared with individual algorithms k-means [8] and Maximum Likelihood Estimator (MLE) [15], as well as APMM [5], WOCCE [8], SMI [13], and BGCM [16] which are state-of-the-art cluster ensemble (selection) methods. This paper reported the empirical results of k-means algorithm as one of the classical clustering methods. Furthermore, as a new alternative in the clustering methods, the empirical results of the proposed method are compared with the MLE, SMI, and BGCM methods. Also, this paper uses the unsupervised version of BGCM method (with the null set of supervision information). For representing the effect of Uniformity on the performance of the final results, it compares with two state-of-the-art metrics in diversity evaluation (APMM and SMI). The last but not least, the experimental results of this paper are compared with the WOCCE as another method in the CES, which uses the independency estimation. All of these algorithms are implemented in the MATLAB R2015a (8.5) by authors[2] in order to generate experimental results. All results are reported by averaging the results of 10 independent runs of the algorithms which are used in the experiment. Also, the number of individual clustering results in the reference set of the ensemble is set as 20 for all of mentioned algorithms in all of experiments on a PC with certain specifications[3].

A. Data Sets

This paper uses three different groups of data sets for generating experimental results; i.e. image based data sets, document based data sets and others. Table I illustrates the properties of these data sets. This paper uses the USPS digits data set, which is a collection of 16×16 gray-scale images of natural handwritten digits and is available from [19]. Furthermore, this paper uses Alzheimer's Diseases Neuroimaging Initiative (ADNI) data set for 202 subjects as another image based real-world data set. This data set contains MRI and PET images from human Brian in two categories (which are shown by C1 and C2 in the Table I

[2]The proposed method is available http://sourceforge.net/projects/myousefnezhad/files/WSCE/

[3]Apple Mac Book Pro, CPU = Intel Core i7 (4*2.4 GHz), RAM = 8GB, OS = OS X 10.10

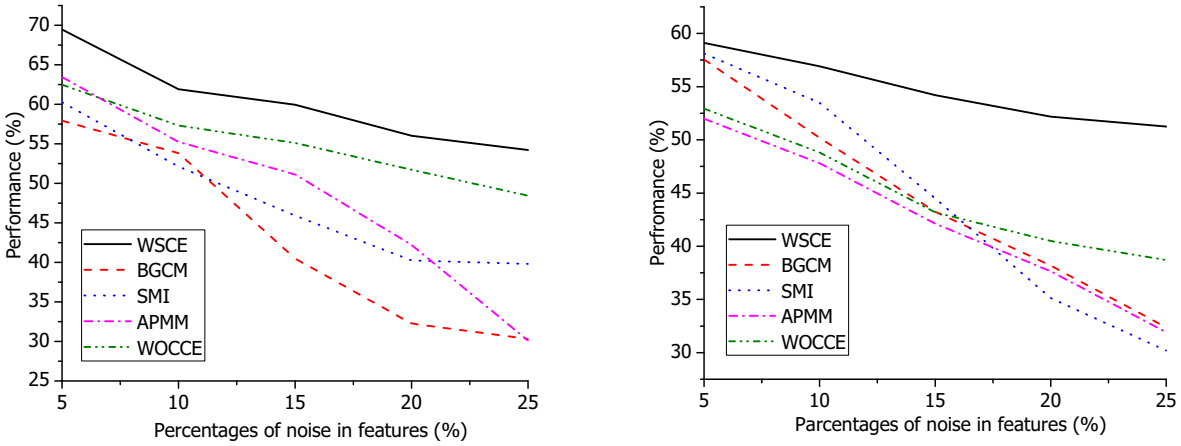
Table II
THE PERFORMANCE OF CLUSTERING ALGORITHMS. FURTHER, THE OPTIONAL FEATURE SELECTION IS NOT USED FOR THE PROPOSED METHOD ($d = 0$).

Data Sets	Spectral	MLE	APMM	WOCCE	SMI	BGCM	WSCE
20 Newsgroups	14.31±2.14	21.89±1.02	28.03±0.87	32.62±0.52	29.14±0.91	40.61±0.83	**52.06±0.17**
ADNI-MRI-C1	39.24±0.21	39.84±0.42	48.01±0.56	48.82±0.37	**50.69±0.69**	45.54±0.99	49.53±0.19
ADNI-MRI-C2	32.72±0.98	26.32±0.67	39.93±0.29	40.22±0.44	38.32±0.41	**42.62±1.04**	41.14±0.71
ADNI-PET-C1	43.71±0.52	37.96±0.87	48.37±0.82	49.19±0.26	49.45±0.62	42.1±0.78	**52.05±0.37**
ADNI-PET-C2	37.27±0.23	37.91±0.83	38.53±0.17	39.43±0.79	41.76±0.47	39.1±1.2	**43.11±0.42**
ADNI-FUL-C1	42.63±0.63	42.62±0.58	47.22±0.93	48.82±0.41	47.93±0.83	48.56±1.26	**49.06±0.36**
ADNI-FUL-C2	39.51±1.19	41.06±0.17	50.09±0.35	49.39±0.63	49.16±0.26	46.91±0.42	**50.11±0.09**
Arcene	58.31±1.22	64.19±0.498	66.28±0.216	65.16±0.32	67.14±0.93	64.23±0.28	**73.34±0.92**
Bala. Scale	49.21±0.87	52.76±0.12	52.65±0.63	54.88±0.61	59.98±0.812	59.62±0.32	**61.64±0.12**
Breast Can.	94.88±1.14	82.65±0.342	96.04±0.88	96.92±0.77	80.87±0.652	99.12±0.62	**99.21±0.43**
Bupa	56.72±1.18	53.98±0.274	55.07±0.28	57.02±0.46	58.49±0.21	53.17±0.21	**60.93±0.09**
CNAE-9	65.32±0.43	77.72±0.591	77.42±0.792	79.2±0.579	74.25±0.614	80.12±0.459	**88.42±0.02**
Galaxy	31.24±0.67	34.25±0.872	33.72±0.36	35.88±0.81	35.21±0.413	36.91±0.17	**39.89±0.82**
Glass	45.78±0.87	50.32±0.42	47.19±0.21	51.82±0.92	54.19±0.144	53.66±0.98	**55.19±0.51**
Half Ring	80.61±1.15	73.91±0.762	80±0.42	87.2±0.14	71.19±0.621	98.37±0.59	**99.92±0.08**
Ionosphere	69.71±0.67	25.67±0.53	70.94±0.13	70.52±0.132	70.87±0.226	73.67±0.341	**76.25±0.28**
Iris	83.45±0.82	89.02±0.61	74.11±0.25	92±0.59	93.79±0.21	**97.29±0.09**	96.53±0.32
Optdigit	54.19±0.45	73.81±0.69	77.1±0.841	77.16±0.21	80.21±0.79	71.56±0.692	**82.82±0.33**
Pendigits	53.94±0.25	59.36±0.31	47.4±0.699	58.68±0.18	63.74±0.37	63.13±0.42	**65.02±0.91**
Reuters-21578	48.78±3.19	52.58±1.92	65.23±0.62	68.85±0.32	62.92±1.02	71.69±0.51	**78.34±0.15**
SA Hart	69.59±0.08	61.69±0.44	70.91±0.42	68.7±0.46	70.05±0.51	**73.92±0.72**	72.8±0.82
Sonar	53.24±0.62	54.93±0.26	54.1±0.91	54.39±0.25	57.64±0.47	52.06±0.873	**61.29±0.11**
Statlog	42.87±0.62	52.35±0.79	54.88±0.528	55.77±0.719	53.73±0.52	55.76±0.591	**57.92±0.26**
USPS	62.67±0.13	59.72±0.62	63.91±0.94	65.21±0.69	68.73±0.66	65.38±1.02	**70.37±0.01**
Wine	73.09±1.38	83.81±0.41	64.6±0.231	71.34±0.542	88.46±0.71	87.34±0.24	**90.44±0.02**
Yeast	32.96±0.71	30.49±0.63	31.06±0.245	32.76±0.268	35.19±0.57	28.12±0.462	**36.92±0.81**

Table I
THE STANDARD DATA SETS

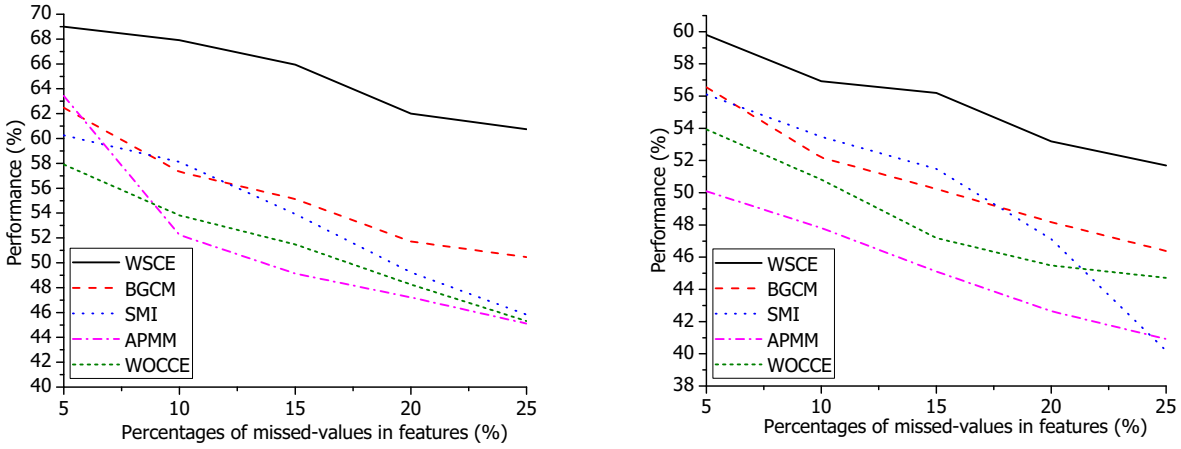
Data Set	Instances	Features	Class
20 Newsgroups	26214	18864	20
ADNI-MRI-C1	202	93	3
ADNI-MRI-C2	202	93	4
ADNI-PET-C1	202	93	3
ADNI-PET-C2	202	93	4
ADNI-FUL-C1	202	186	3
ADNI-FUL-C2	202	186	4
Arcene	900	10000	2
Bala. Scale	625	4	3
Brea. Cancer	286	9	2
Bupa	345	6	2
CNAE-9	1080	857	9
Galaxy	323	4	7
Glass	214	10	6
Half Ring	400	2	2
Ionosphere	351	34	2
Iris	150	4	3
Optdigit	5620	62	10
Pendigits	10992	16	10
Reuters-21578	9108	5	10
SA Hart	462	9	2
Sonar	208	60	2
Statlog	6435	36	7
USPS	9298	256	10
Wine	178	13	2
Yeast	1484	8	10

and II) for recognizing the Alzheimer diseases. In the first category, this data set partitions subjects to three groups of Health Control (HC), Mild Cognitive Impairment (MCI), and Alzheimer's Diseases (AD). In the second category, there are four groups because the MCI will be partitioned to high and low risk groups (HMCI/LMCI). This paper uses all possible forms of this data set by using only MRI features, only PET features and all of MRI and PET features (FUL) in each of two categorize. More information about ADNI-202 is available in [20]. As a document based data set, the 20 Newsgroups is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. Some of the newsgroups are very closely related to each other, while others are highly unrelated. It has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering. Moreover, the Reuters-21578 is one of the most widely used test collections for text classification research. This data set was collected and labeled by Carnegie Group, Inc. and Reuters, Ltd. We use the 10 largest classes of this data set. The rest of standard data sets are from UCI [21]. The chosen data sets have diversity in their numbers of clusters, features, and samples. Further, their features are normalized to a mean of 0 and variance of 1, i.e. N (0, 1).



(a) Arcene (b) CNAE-9

Figure 2. The effect of noisy data sets on the performance.



(a) Arcene (b) CNAE-9

Figure 3. The effect of missed-values on the performance.

B. Performance analysis

In this section the performance (accuracy metric [8]) of proposed method will be analyzed. In other words, 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 [8]. The results of the proposed method are compared with individual algorithms Spectral clustering[17] and MLE [15], as well as APMM [5], WOCCE [8], SMI [13], and BGCM [16] which are state-of-the-art cluster ensemble (selection) methods. The main reason for comparing the proposed method with Spectral clustering is to show the effect of TKSC framework on the performance of the final results. Furthermore, as a new alternative in the graph based clustering methods, the empirical results of WSCE are compared with the MLE and BGCM methods. This paper uses the unsupervised version of BGCM method (with the null set of supervision information). For representing the effect of Normalized Modularity on the performance of the final results, it compares with three state-of-the-art metrics in diversity evaluation (A3, APMM and SMI), which are based on Shannons entropy. This paper doesn't use optional feature selection in this section ($d = 0$). The experimental results are given in Table II. In this table, the best result which is achieved for each data set is highlighted in bold. As depicted in this table, although individual clustering algorithms (Spectral and MLE) have shown acceptable performance in some data sets, they cannot recognize true patterns in all of them. As mentioned earlier in this paper, in order to solve the clustering problem, each individual algorithm considers a special perspective of a data set which is based on its objective function. The achieved results of individual clustering algorithms, which are depicted in

Table II are good evidence for this claim. Furthermore, the results generated by APMM, SMI, and WOCCE show the effect of the aggregation method on improving accuracy in the final results. According to Table II, BGCM and the proposed algorithm (WSCE) have generated better results in comparison with other individual and ensemble algorithms. Even though the proposed method was outperformed by a number of algorithms in four data sets (Iris, SA Hart, and ADNI-MRI-C1/C2), the majority of the results demonstrate the superior accuracy of the proposed method in comparison with other algorithms. In addition, the difference between the performance of proposed method and the best result in those four data sets is lower that 2%.

C. Noise and missed-values analysis

The effect of noise and missed-values on the performance of clustering algorithms will be discussed in this section. The optional feature selection for the proposed method doesn't use in this section (d = 0). In Figure 2, the effect of noise in the features of data sets will be analyzed on the performance of proposed method. This figure represents the performance of the WSCE, WOCCE, BGCM, SMI, and APMM on the noisy data sets. In this experiment, some features of Arcene and CNAE-9 data sets are randomly changed. This figure shows that proposed method generates more stable results because the Normalized Modularity provides a robust diversity evaluation for selecting most stable individual results. As mentioned before, Shannon's entropy uses the logarithm of probability of individual results for evaluating the diversity but there is no mathematical prove that all real-world data sets have logarithmic behavior. This experiment is the best evidence for this claim. Figure 3 demonstrates the analysis for the effect of missed-values in the data sets on the performance of clustering algorithms. This figure illustrates the performance of the WSCE, WOCCE, BGCM, SMI, and APMM on the data sets with missed-values. In this experiment, some values of attributes of Arcene and CNAE-9 data sets are randomly missed (set null). As you can see in this Figure, the proposed method and BGCM generate more stable results. This is a new advantage of our proposed method in comparison other non-graph based methods. Since, our proposed method uses the TKSC algorithms for generating Partitional and Modular results, it can significantly handle the miss values. In other words, as a local error in the individual results, a missed-value just can destroy an edge in our Modular result, which can be recognized by comparing Modular result with Partitional result in the diversity evaluation by using the NM metric. That is another reason for exploiting the proposed framework in the clustering problems.

D. Parameter analysis

In this section the performance of the proposed method will be analyzed by using the optional features selection

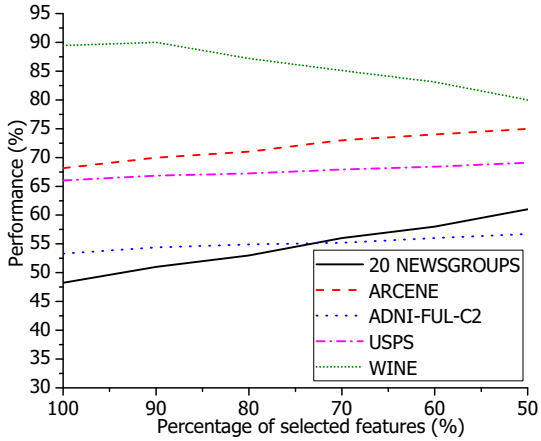


Figure 4. The effect of optional features selection on the performance of proposed method.

(d parameter). This paper employs various data sets, i.e. two low dimension data sets (Wine, Glass), two high-dimension data sets (20 Newsgroups, Arcene), and two middle-dimension and also image based data sets (USPS, ADNI) for analyzing the performance of proposed method. Figure 4 illustrates the relationship between the performance of the proposed method based on the percentage of selected features in different data sets. The vertical axis refers to the performance while the horizontal axis refers to the percentage of selected feature in each data set. As you can see in this figure, the optional feature selection can significantly increase the performance of final results on high-dimensional data sets; and also it can dramatically decrease the performance on low-dimensional data sets. Further, it is not more effective on the middle-dimension data sets. This paper offers that the optional features selection will be used only for high-dimensional data sets for handling features-sparsity.

V. Conclusion

There are two challenges in Cluster Ensemble Selection (CES); i.e. proposing a robust consensus metric(s) for diversity evaluation and estimating optimum parameters in the thresholding procedure for selecting the evaluated results. This paper introduces a novel solution for solving mentioned challenges. By employing some concepts from community detection arena and graph based clustering, this paper proposes a novel framework for clustering problems, which is called Weighted Spectral Cluster Ensemble (WSCE). Under this framework, a new version of spectral clustering, which is called Two Kernels Spectral Clustering (TKSC), is used for generating graphs based individual clustering results; i.e. Partitional result and Modular result. Instead of entropy based methods in the traditional CES, this paper introduces Normalized Modularity (NM), which is a redefined version of modularity in the community detection arena for general

clustering problems. The NM is used on the transformed graph representation of individual clustering results for providing an effective diversity estimation. Moreover, this paper introduces a new solution for combining the evaluated individual clustering results without the procedure of thresholding, which is called Weighted Evidence Accumulation Clustering (WEAC). While the weight of each individual result in WEAC can be estimated with different metrics, the NM was used in this paper. To validate the effectiveness of the proposed approach, an extensive experimental study is performed by comparing with individual clustering methods as well as cluster ensemble (selection) methods on a large number of data sets. Results clearly show the superiority of our approach on both normal data sets and those with noise or missing values. In the future, we plan to develop a new version of normalized modularity for estimating the diversity of Partitional results, directly.

ACKNOWLEDGMENT

We thank Dr. Sheng-Jun Huang for his helpful suggestions, and the anonymous reviewers for comments. This work was supported in part by the National Natural Science Foundation of China (61422204 and 61473149), Jiangsu Natural Science Foundation for Distinguished Young Scholar (BK20130034) and NUAA Fundamental Research Funds (NE2013105).

REFERENCES

[1] A. Strehl and J. Ghosh, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2002.

[2] A. Fred and A. Lourenco, "Cluster ensemble methods: from single clusterings to combined solutions," *Computer Intelligence*, vol. 126, pp. 3–30, 2008.

[3] X. Fern and W. Lin, "Cluster ensemble selection," in *SIAM International Conference on Data Mining (SDM'08)*, 24-26 April 2008, pp. 128–141.

[4] J. Azimi and X. Fern, "Adaptive cluster ensemble selection," in *21th International joint conference on artificial intelligence (IJCAI-09)*, 11-17 July 2009, pp. 992–997.

[5] H. Alizadeh, B. Minaei-Bidgoli, and H. Parvin, "Cluster ensemble selection based on a new cluster stability measure," *Intelligence Data Analysis (IDA)*, vol. 18, no. 3, pp. 389–40, 2014.

[6] J. Jia, X. Xiao, and B. Liu, "Similarity-based spectral clustering ensemble selection," in *9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 29-31 May 2012, pp. 1071–1074.

[7] M. Yousefnezhad, H. Alizadeh, and B. Minaei-Bidgoli, "New cluster ensemble selection method based on diversity and independent metrics," in *5th Conference on Information and Knowledge Technology (IKT'13)*, 22-24 May 2013.

[8] H. Alizadeh, M. Yousefnezhad, and B. Minaei-Bidgoli, "Wisdom of crowds cluster ensemble," *Intelligent Data Analysis (IDA)*, vol. 19, no. 3, 2015.

[9] A. Clauset, M. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, no. 066111, 2004.

[10] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 23, pp. 8577–8696, 2006.

[11] X. Zhang, L. Zhao, L. Zong, and X. Liu, "Multi-view clustering via multi-manifold regularized nonnegative matrix factorization," in *IEEE International Conference on Data Mining series (ICDM'14)*, 15–17 December 2014.

[12] A. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 835–850, 2005.

[13] S. Romano, J. Bailey, N. X. Vinh, and K. Verspoor, "Standardized mutual information for clustering comparisons: One step further in adjustment for chance," in *31st International Conference on Machine Learning (ICML14)*, 21-26 June 2014, pp. 1143–1151.

[14] C.-T. Kuo, P. Walker, O. Carmichael, and I. Davidson, "Spectral clustering for medical imaging," in *IEEE International Conference on Data Mining series (ICDM'14)*, 15–17 December 2014.

[15] Y. Chen, S. H. Lim, and H. Xu, "Weighted graph clustering with non-uniform uncertaintiese," in *31st International Conference on Machine Learning (ICML14)*, 21-26 June 2014, pp. 1566–1574.

[16] J. Gao, F. Liang, W. Fan, Y. Sun, and J. Han, "A graph-based consensus maximization approach for combining multiple supervised and unsupervised models," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 1, pp. 15–2, 2013.

[17] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems 14 (NIPS'01)*, 2001, pp. 849–856.

[18] D. Yan, L. Huang, and M. I. Jordan, "Fast approximate spectral clustering," in *15th ACM Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2009.

[19] S. Roweis. (1998) The world-famous courant institute of mathematical sciences, computer science department, new york university. [Online]. Available: http://cs.nyu.edu/roweis/data.html

[20] C. Zu and D. Zhang, "Label-alignment-based multi-task feature selection for multimodal classification of brain disease," in *4th NIPS Workshop on Machine Learning and Interpretation in Neuroimaging (MLINI'14)*, 13 December 2014.

[21] C. B. D. J. Newman, S. Hettich, and C. Merz. (1998) Uci repository of machine learning databases. [Online]. Available: http://www.ics.uci.edu/mlearn/MLSummary.html