A kernel-induced weighted object-cluster association-based ensemble method for educational data clustering

Chau Thi Ngoc Vo and Phung Hua Nguyen

Ho Chi Minh City University of Technology, Vietnam National University, Ho Chi Minh City, Vietnam

ABSTRACT
Exploiting single models for a better ensemble one is popular for a classification or clustering task. For educational data clustering, our work proposes a kernel-induced weighted object-cluster association ensemble method, named kWOCA, to discover the inherent structure of the data and return the student clusters of higher quality. kWOCA is an advanced version that can automatically determine the number of desired clusters based on Bayesian Information Criterion. It also conducts consensus clustering in the feature space for non-linearly separated clusters where the discrimination between the objects is captured from the base clusterings. Besides, it encodes the differences between the clusters in each base clustering and those between the base clusterings in the ensemble when making a synthesis of the base clusterings. A kernel-induced weighted object-cluster association matrix is defined to store such rich information. Using this matrix, kWOCA outperforms the existing clustering ensemble methods. Experimental results on the real educational data sets and the benchmark Iris data set show the better effectiveness of kWOCA with higher Normalized Mutual Information values. As a result, groups of the most similar students based on their study performance can be discovered better. These resulting student clusters can be then further analyzed for academic affairs.

1. Introduction

Educational data mining is realized to be an interesting and significant research area for knowledge discovery in educational data. Its achievements are utilized for enhancing learning and teaching activities as well as administrative processes in an educational organization. Among the educational data mining tasks, educational data clustering is well known for its capability to discover inherent structures in the data in terms of clusters of the most similar objects. Many existing works have been proposed with many expected results and different clustering algorithms. Some of them are listed as: Adjei, Ostrow, Erickson, and Heffernan (2017) discovered similar student behaviours, Bresfelean, Bresfelean, and Ghisoiu (2008) aimed at similar profiles, and (Jovanovic, Vukicevic, Milovanovic, & Minovic, 2008) aimed at similar profiles, and (Jovanovic, Vukicevic, Milovanovic, & Minovic,
2012; Kerr & Chung, 2012; Vo & Nguyen, 2018) focussed on similar performance. As for clustering algorithms, many popular existing algorithms: $k$-means, fuzzy $c$-means, spectral clustering, Self-Organizing Map, etc. are available nowadays. Among them, $k$-means proposed in MacQueen (1967) has been widely-used in many works (Adjei et al., 2017; Bresfelean et al., 2008; Campagni, Merlini, & Verri, 2014; Jovanovic et al., 2012), while the Partitional Segmentation algorithm in Jayabal and Ramanathan (2014) and the FANNY and AGNES algorithms in Kerr and Chung (2012). Although efficient and effective in some cases, $k$-means is influenced by the existence of noises and only its intra-cluster compactness is taken into account. Consequently, it can achieve the quality of the clusters to some certain extent due to its Gaussian spherical assumption for the resulting clusters. In addition, it is noted that except (Vo & Nguyen, 2018) with the proposed WOCA ensemble method, none of the existing educational data clustering works has considered the resulting clusters of higher quality with any ensemble method. As one of the first works advancing the educational data clustering task with an ensemble method for better clusters, more robustness, and knowledge reuse (Topchy, Jain, & Punch, 2005; Zhou, 2012), our work concentrates on the development of an extended version of WOCA in this paper.

Ensemble clustering has received great attentions in the data mining research area. However, encoding the details of both object-cluster associations and differences between the clusters has not yet been examined thoroughly in the existing works. Using embedding-based approach, Franek and Jiang (2014) obtained median consensus clustering with no predefined $k$ desired clusters. Unfortunately, no detail of each base clustering was considered when embedding base clusterings in a vector space and treating each of them as a point in that space. As for (Huang, Wang, & Lai, 2017; Ren, Domeniconi, Zhang, & Yu, 2016; Zhong, Yue, Zhang, & Lei, 2015) based on a co-association matrix that captured the co-occurrence of the objects in the same cluster, more information has been added into each value in the matrix or into each object in terms of weights. Detailed review about these works was given in Vo and Nguyen (2018). By contrast, our work approaches the task with an object-cluster association matrix. Therefore, the main difference between them and our work is that our method makes the contribution of each cluster in a base clustering in the consensus clustering process more visible and clearer. Our method also combines base clusterings in such a way that the objects can get more discriminated and better clustered. A higher dimensional feature space is of our interest so that the separation between the objects and that between the clusters can be well determined. Similar to ours, Liu, Wu, Liu, Tao, and Fu (2017), Pattanodom, Iam-On, and Boongoen (2016), and Wu, Liu, Xiong, Cao, and Chen (2015) used an object-cluster association matrix, based on $k$-means as reviewed in Vo and Nguyen (2018). For Liu et al. (2017), weighted $k$-means was considered to avoid the complexity of spectral ensemble clustering with spectral clustering (Ng, Jordan, & Weiss, 2002). In comparison, these works exploited only a binary object-cluster association matrix while ours proposes a kernel-induced weighted object-cluster association matrix and does consensus clustering on this new ensemble data matrix for better final clusters.

In short, in this paper, we propose a kernel-induced weighted object-cluster association-based ensemble method, named $k$WOCA, for educational data clustering. The expected output is the high-quality clusters of the most similar students based on their similar study performance. The number of desired clusters in consensus clustering is automatically determined with the largest Bayesian Information Criterion (BIC) value. In addition, $k$WOCA is based on its corresponding kernel-induced weighted object-cluster
association matrix, defined in a higher dimensional feature space, using the Gaussian kernel function. Besides, the differences between the clusters in a base clustering and those between the base clusterings in the ensemble model are examined with their BIC values. When we embed more aspects of the objects, then of their clusters, and then of their base clusterings into the matrix, the matrix gets richer and thus becomes a better input data set for the next clustering phase.

Through the experimental results on the real educational data sets and the benchmarking popular Iris data set, our resulting clusters are of higher quality with better Normalized Mutual Information values. In particular, compared to k-means, kWOCA can derive the final clustering composed of the clusters of higher quality. In addition, kWOCA outperforms the existing methods based on either a co-association matrix or a binary object-cluster association one. It also outperforms its previous version, WOCA. These achievements confirm the effective design of kWOCA for ensemble clustering. Therefore, our educational data clustering task can be handled better and provide many various groups of the most similar students. The group of the in-trouble students based on their study performance can be identified and supported well.

The rest of our paper is organized as follows. Section 2 defines our educational data clustering task. In Section 3, we propose a kernel-induced weighted object-cluster association-based ensemble method, kWOCA. Its descriptions, pseudo codes, and analysis are also given. Section 4 details an empirical evaluation of our kWOCA method in comparison with some existing ones. In Section 5, several concluding remarks are presented along with our future works.

2. A study performance-based student clustering task

A data clustering task is popular among data mining tasks to discover groups of similar objects in a given data set where only data characteristics of each object are available with no predefined class. In the educational domain, this task is thus used to group similar students together based on different characteristics observed for each student. In our work, each student is characterized by his/her study performance with respect to the enrolled program. This study performance is reflected via the grades the student obtained in the courses. For those courses not available, zeros were used to represent no study performance of the student. Therefore, students with similar study performance can be examined at the program level. Further cluster analysis can be explored to discover the students who might not complete the program in time.

In a computational form, a vector space model is used. A $p$-dimension vector is defined for each student where $p$ is the dimensionality of the space, corresponding to the number of subjects required by the program. Each element at a dimension is a grade of a subject in the range $[0, 10]$. In this space, $n$ vectors corresponding to $n$ students are examined for the clustering task to find the student groups each of which contains the most similar students based on their study performance.

Given the aforementioned context, our clustering task is formally stated as follows.

The input is a data set $D$ of $n$ object vectors in a $p$-dimensional data space:

$$ D = \{X_1, X_2, \ldots, X_n\} \quad (1) $$

where $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $x_{id} \in [0, 10]$ for $\forall i = 1 \ldots n, d = 1 \ldots p$. 
The output is a clustering $C$ that captures the inherent structure of the data set $D$ in $k$ clusters where $k$ is the number of clusters determined in the clustering process. Each cluster $C_j$ contains the $n_j$ most similar vectors for $j = 1 \ldots k$.

$$C = \{C_1, C_2, \ldots, C_k\}$$

where $C_j = \{X_{j,1}, X_{j,2}, \ldots, X_{j,n_j}\}$ and $D = C_1 \cup \cdots \cup C_k$ for $\forall j = 1 \ldots k$, $1 \leq n_j \leq n$.

In this work, an ensemble method is proposed to obtain the output with the input described above. The output is expected to be better than that from a single method.

### The proposed ensemble method

In this section, we propose a kernel-induced weighted object-cluster association-based ensemble method for educational data clustering. We name the proposed method kWOCA. It is an extended version of WOCA in Vo and Nguyen (2018) and thus, inherits all the characteristics of WOCA. In particular, kWOCA also follows the pair-wise co-occurrence-based approach. kWOCA employs $k$-means for its efficiency. kWOCA exploits the differences between the objects in each base clustering like WOCA. However, in its extension to WOCA, it gives more information about the differences between the clusters in each base clustering and the differences between the base clusterings. It is expected that kWOCA outperforms WOCA and the other methods.

#### 3.1. kWOCA: from base clusterings to a kernel-induced weighted object-cluster association matrix

Similar to WOCA in Vo and Nguyen (2018), kWOCA is elaborated with three main phases ((1). Base clusterings construction, (2). Ensemble data matrix construction, (3). Consensus clustering) using the following notations for its input, defined in Equation (1):

$D = \{X_1, X_2, \ldots, X_n\}$ for an input data set where $X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})$, $\forall i = 1 \ldots n$, $n$ for the number of objects, $p$ for the number of attributes, $e$ for the ensemble size, i.e. the number of base clusterings. Its output is a final clustering with $k$ clusters where $k$ is automatically determined by the Bayesian Information Criterion (BIC)-based model selection.

**Phase (1). Base clusterings construction**

This phase corresponds to the first phase of WOCA in Vo and Nguyen (2018). The difference between them is an automatic determination of the number of desired clusters, $k$, for each base clustering and consensus clustering. The BIC-based model selection well-known with $k$-means proposed in Pelleg and Moore (2000) is used. It is a suitable choice as our base clustering algorithm is $k$-means, also with the underlying spherical Gaussian assumption. According to the formulas in Pelleg and Moore (2000), we obtain a BIC value for each clustering generated with $k = 2 \ldots (\sqrt{n} + 1)$. The selected value $k$ is associated with the maximum BIC value, showing the maximum log-likelihood of the objects belonging to their clustering.

Using the selected value $k$, $k$-means with random initialization is executed on the original data set to generate $e$ base clusterings. Each base clustering is obtained by minimizing
the objective function until their resulting clusters are unchanged:

$$F_r = \sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_{ij} d^2(X_i, C_{rj}), \quad \forall \ r = 1 \ldots e$$

where $\gamma_{ij}$ is the membership of $X_i$ to the cluster whose centre (representative) is $C_{rj}$: 1 if a member; otherwise, 0; and $d(X_i, C_{rj})$ is a distance between $X_i$ and the centre $C_{rj}$.

**Phase (2). Ensemble data matrix construction**

In this phase, a kernel-induced weighted object-cluster association (kWOCA) matrix, named kWOCA_matrix, is defined with $n$ rows and $ek$ columns. Each row represents an object in the ensemble space. Each column represents a cluster in a base clustering, playing a role of a dimension of the ensemble space. It has the same size as WOCA_matrix in Vo and Nguyen (2018). However, it can reflect more characteristics of objects and their clusters from the base clusterings.

In kWOCA_matrix, a cell value at position $(i,l)$ shows the closeness of the object $X_i$ to the representative of the corresponding cluster $C_l$, which is the association between an object and a cluster. This association is weighted for more details rather than a binary association. Different from the weighting scheme in WOCA_matrix, the weighting scheme in kWOCA_matrix is proposed more comprehensively from all the aspects of an object and its association with all the clusters in base clusterings.

First, to differentiate between the objects in each cluster in a base clustering, a kernel-induced Euclidean distance ratio is defined in the similar manner that we defined an association weight in WOCA_matrix in Vo and Nguyen (2018). Instead of the original data space, a higher-dimensional feature space using an implicit transformation with the Gaussian kernel function is considered. The Euclidean distance, showing the closeness of an object and each cluster in the original data space, now becomes a kernel-induced Euclidean distance, showing the closeness of an object and each cluster in the feature space. Their separation is expanded in this space and thus, the overlapping between the clusters is reduced. As a result, this characteristic is expected to enhance the quality of the clusters discovered in overlapping data sets. However, its calculation increases a bit our computational cost. Their details are given below:

The Euclidean distance between an object $X_i$ and the centre of a cluster $C_{rj}$ in a base clustering $\pi_r$ is $d(X_i, C_{rj})$. Their kernel-induced Euclidean distance in a higher dimensional feature space is $kd(X_i, C_{rj})$:

$$kd(X_i, C_{rj}) = 2 - 2K(X_i, C_{rj})$$

where $K(\cdot)$ is a Gaussian kernel function with a bandwidth $\sigma$:

$$K(X_i, C_{rj}) = e^{-\frac{(d(X_i, C_{rj})^2)}{2\sigma^2}}$$

In our work, $\sigma$ is derived as a ratio $r_\sigma$ to the variance var of $D$ from each dimension as follows. This calculation can allow users to easily determine this ratio as how much the differences between the objects over all the dimensions will be retained and
considered for a nonlinear transformation from the original data space to a feature space.

$$\sigma = r_\alpha \times \text{var} = r_\alpha \sum_{i=1}^{p} \frac{1}{n} \left( \sum_{i=1}^{n} (x_i^j)^2 - \frac{1}{n} \left( \sum_{i=1}^{n} nx_i^j \right)^2 \right)$$  \hspace{1cm} (6)

The maximum distance between an object belonging to a cluster \(C_{r,j}\) and its centre in a base clustering \(\pi_r\) is \(\text{Radius}_{r,j}\):

$$\text{Radius}_{r,j} = \max_{i=1...n} \gamma_{i,j} kd(X_i, C_{r,j})$$  \hspace{1cm} (7)

Where \(\gamma_{i,j}\) is the membership of \(X_i\) to the cluster \(C_{r,j}\) of a base clustering \(\pi_r\).

$$\gamma_{i,j} = \begin{cases} 1 & \text{ } X_i \in C_{r,j} \\ 0 & \text{Otherwise} \end{cases}$$  \hspace{1cm} (8)

A kernel-induced Euclidean distance ratio \(r_1\) between an object \(X_i\) and the centre of a cluster \(C_{r,j}\) in a base clustering \(\pi_r\) is then defined as follows:

$$r_1 = \frac{kd(X_i, C_{r,j})}{\text{Radius}_{r,j}}$$  \hspace{1cm} (9)

Second, to differentiate between the clusters in each base clustering and between the base clusterings in the ensemble, a level-wise BIC-based ratio is defined. Once obtaining a base clustering \(\pi_r\) for the ensemble model, we derive a \(\text{BIC}_{r,j}\) value for each cluster \(C_{r,j}\) and another one \(\text{BIC}_r\) for that base clustering, which is defined as total sum of the \(\text{BIC}_{r,j}\) values of all the clusters in that base clustering. The BIC value for the entire ensemble is defined as the total sum of the \(\text{BIC}_r\) values of all the base clusterings. The larger BIC value shows the better model. Thus, at the base clustering level, we use \(\text{BIC}_{r,j}/\text{BIC}_r\) to represent better/worse clusters in a base clustering and at the consensus clustering level, we use \(\text{BIC}_r/\text{BIC}\) to represent better/worse base clusterings in the ensemble. For simplicity, a product is used for their combination. A level-wise BIC-based ratio \(r_2\) for a cluster \(C_{r,j}\) in a base clustering \(\pi_r\) is defined:

$$r_2 = \frac{\text{BIC}_{r,j}/\text{BIC}_r}{\text{BIC}} = \frac{\text{BIC}_{r,j}}{\text{BIC}}$$  \hspace{1cm} (10)

Putting them altogether, an association between an object \(X_i\) and a cluster \(C_{r,j}\) in a base clustering \(\pi_r\) at cell \((i, (r - 1) \times k + j)\) is weighted as follows:

$$k\text{WOCA\_matrix}[i][j][(r - 1) \times k + j] = r_1 \times r_2 = \frac{kd(X_i, C_{r,j}) \text{BIC}_{r,j}}{\text{Radius}_{r,j} \text{BIC}}$$  \hspace{1cm} (11)

Like \(\text{WOCA\_matrix}\), \(k\text{WOCA\_matrix}\) can capture both membership and membership degree of an object to a cluster. However, it captures those in a higher dimensional feature space. Besides, it can consider the contribution of each cluster to consensus clustering differently according to its quality in a base clustering and the quality of its base clustering in the ensemble. In short, \(k\text{WOCA\_matrix}\) is more informative than \(\text{WOCA\_matrix}\) and expected to be a better input for consensus clustering.
**Phase (3). Consensus clustering**

Using kWOCA_matrix output from the second phase, consensus clustering is conducted in this phase to derive the final clustering. In our work, k-means is used with the number of desired clusters, $k$, determined in the first phase. In the clustering process, the following objective function is minimized until the stability of the $k$ resulting clusters is reached for convergence.

$$F = \sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_{ij} d^2(X_i, C_j)$$

(12)

where $\gamma_{ij}$ is the membership of the object $X_i$ to the cluster $C_j$: 1 if a member; otherwise, 0; and $d(X_i, C_j)$ is a distance between $X_i$ and $C_j$ in their current feature space.

The difference between $F_r$ for $r = 1 \ldots e$ and $F$ is the computation of the distance: $d(X_i, C_{rj})$ vs. $d(X_i, C_j)$. Each distance in $F_r$ is computed in the original space while $F$'s distance in the ensemble space after an implicit non-linear transformation. Therefore, non-spherical clusters can be discovered in our work although the clustering process is based on the Gaussian spherical assumption with BIC values and k-means. Indeed, kWOCA_matrix has encoded the differences between the objects and the clusters in a feature space and linear separations between the clusters in a feature space are in fact non-linear ones in its original space.

In Figure 1, pseudo code of the proposed kWOCA ensemble method is detailed.

**3.2. Characteristics of the proposed method**

Generally speaking, our kWOCA method inherits the strengths of the WOCA method in Vo and Nguyen (2018) and deploys more characteristics in the input data and the generated base clusterings for consensus clustering. As k-means is a base clustering algorithm in our ensemble model, the simplicity and efficiency of k-means can be preserved in the first and third phases of kWOCA. Although other existing clustering algorithms can be used as a base clustering algorithm, k-means is often an appropriate choice. In the following, kWOCA is analyzed in terms of space and time. It is also compared with the existing methods based on the co-association matrix.

For space, $O(nek)$ is the size of the kernel-induced weighted object-cluster association matrix while $O(n^2)$ is the size of the co-association matrix. Our space complexity can be $O(en\sqrt{n})$ if the BIC-based model selection scheme returns $(\sqrt{n} + 1)$ for the number of clusters. Even in that case, when the ensemble size $e$ is fixed and often smaller than the number of the objects $n$, the cost with our matrix is lower than that with the co-association matrix. With a smaller ensemble data matrix, the computational cost on such a resulting matrix will then decrease accordingly.

For time, it is supposed that kWOCA_matrix is used and any base clustering algorithm can be employed to obtain a final consensus clustering. However, we can have a much lower time complexity with k-means.

For phase (1), time complexity to determine the number $k$ of desired clusters based on BIC values is $O(n^2t)$ when $k$ is from 2 to $(\sqrt{n} + 1)$, $t$ is the number of iterations, $n$ is the number of objects in the $p$-dimensional space for $p \ll n$. Time complexity to generate $e$ base clusterings is $O(ent)$ when it is supposed that the maximum value $(\sqrt{n} + 1)$ is selected for $k$. Therefore, the final time complexity in this phase is...
$$T_1 = O(n^2t) + O(\text{ent}) \simeq O(n^2)$$ for phase (2), time complexity is $$T_2 = O(nek)$$ stemming from the traversing on kWOCA_matrix corresponding to its size, which is $$O(nk)$$ for space and given $$k$$ is the number of desired clustered returned by phase (1). For phase (3), time complexity is $$T_3 = O(nkt')$$ for consensus clustering. Putting them altogether, time complexity is $$T_1 + T_2 + T_3 = O(n^2) + O(nek) + O(nkt') \simeq O(n^2)$$ in total when the number of desired clusters is automatically determined. For the data sets with the predefined number $$k$$ of desired clusters, it turns nearly linear with $$O(n)$$. Such a reasonable time complexity enables us to consider $$k$$-means as a base clustering algorithm in our method and the others in the existing works (Liu et al., 2017; Ren et al., 2016; Topchy et al., 2005; Vo & Nguyen, 2018; Wu et al., 2015).

From the user's perspectives, kWOCA has two parameters: $$e$$ for the ensemble size and $$r_\sigma$$ for the ratio to the data set's variance for the bandwidth $$\sigma$$ of the Gaussian kernel.

---

**Algorithm**: kWOCA ensemble

**Input:**
- Data set: $$D = (X_1, X_2, \ldots, X_n)$$
- Number of base clusterings (i.e. ensemble size): $$e$$
- Ratio to $$D$$'s variance for the bandwidth $$\sigma$$ of the Gaussian kernel function: $$r_\sigma$$

**Output:**
- Final clustering $$\pi^* = C = \{C_1, C_2, \ldots, C_k\}$$ where $$k$$ is the number of clusters determined in consensus clustering

**Process:**
// Phase (1). Base clusterings construction
// Obtain $$k$$, the number of clusters
(1). For $$i = 2\ldots\sqrt{n} + 1$$
  // BIC is based on (Wu et al., 2015)'s formulas.
  BIC_i = getBIC(k-means(D, i));
(2). $$k = \text{argmax}_{i=2\ldots\sqrt{n}+1} \text{BIC}_i$$;
// Obtain the base clusterings
(3). For $$r = 1\ldots e$$
(4). Each base clustering $$\pi_r = k$$-means($$D$$, $$k$$);
// Phase (2). Ensemble data matrix construction
(5). kWOCA_matrix = a kernel-induced BIC-based weighted object-cluster association matrix of the kWOCA method;
(6). For each base clustering $$\pi_r$$ where $$r = 1\ldots e$$
(7). For each cluster $$C_{r, j}$$ in $$\pi_r$$ where $$j = 1\ldots k$$
(8). Get $$\text{Radius}_{r, j}$$ according to (7);
(9). For each object $$X_i$$ in $$D$$ where $$i = 1\ldots n$$
(10). For each base clustering $$\pi_r$$ where $$r = 1\ldots e$$
(11). For each cluster $$C_{r, j}$$ in $$\pi_r$$ where $$j = 1\ldots k$$
(12). Get $$\text{kWOCA_matrix}[i][(r-1)k + j]$$ according to (11);
// Phase (3). Consensus clustering
(13). Final clustering $$\pi^* \leftarrow k$$-means(kWOCA_matrix, $$k$$);
(14). Return $$\pi^*$$;

**Figure 1.** Pseudo code of the proposed kWOCA method.
function. The first parameter can be determined according to the need and available requirement of the application. It is a widely-used parameter in almost all the ensemble methods. The second parameter can be determined according to the data characteristics in order to make a non-linear space transformation. If users have prior knowledge about their data spaces, they can easily set a value to this parameter. Otherwise, a grid search scheme or any other hyperparameter tuning methods can be applied to determine an appropriate value for this parameter. It is believed that these two parameters of kWOCA are well studied and don’t place any burden on its users.

In short, it is confirmed that kWOCA is among the most recent works considering the difference between the objects in a single cluster of each base clustering via their representatives and thus, better distinguishing between the objects in consensus clustering. Its novelty is also with the kernel-induced level-wise BIC-based weighted object-cluster association. Such associations can be captured more comprehensively in detail. As a result, consensus clustering can then pay attention more to the contributions of the good base clusterings and explore more the differences between the objects with respect to each cluster in each base clustering. Our final clustering ensemble model is expected to have higher quality than those with the other methods.

4. An empirical evaluation

To evaluate the new features for the proposed kWOCA method, an empirical study is conducted on the same data sets as used in Vo and Nguyen (2018): ‘Year 4 CE’ and ‘Year 4 CS’ from Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology, Vietnam National University – Ho Chi Minh City (Academic Affairs Office, 2017) and Iris from Iris (2017). ‘Year 4 CE’ is a data set that contains the study results of the fourth-year undergraduate students enrolled in 2008–2009 with the Computer Engineering program while ‘Year 4 CS’ in 2005–2008 with the Computer Science program. Each student is characterized by 43 subjects typically required for each program. Once performing the clustering task on these two data sets, we expect to group the similar students based on their study performance. The students facing difficulties in their study can belong to the same group while the students with success in other groups. These two data sets are real and private. Therefore, Iris, which is a popular benchmarking data set with 150 instances, 4 attributes and 3 classes, is used to demonstrate how well the clustering task can be solved. Their details are given in Table 1.

Different from Vo and Nguyen (2018), we use the predefined number of clusters, 3, for Iris while determining the optimal number of clusters for each real educational data set. That is a kernel-induced weighted object-cluster association matrix is straightforwardly defined for Iris with 3 clusters for each base clustering, while such a matrix is

| Data set | Year 4 CE | Year 4 CS | Iris |
|----------|-----------|-----------|------|
| Program  | Computer Engineering | Computer Science | (Benchmarking) |
| Entrance Year | 2008–2009 | 2005–2008 | – |
| Number of Instances | 186 | 1317 | 150 |
| Number of Attributes | 43 | 43 | 3 |
| Number of Classes | 3 | 3 | 3 |
| Number of Clusters | Varying | Varying | 3 |
defined for educational data sets with the non-predefined number \( k \) of clusters for each base clustering after BIC-based model selection is examined. An optimal value for \( k \) is selected in the range \([2, \sqrt{n} + 1]\), where \( n \) is the number of instances. Besides, similar to Vo and Nguyen (2018), the number of base clusterings, i.e. the ensemble size \( e \), varies from 10 to 80 with a gap of 10 for each data set. In addition, random initialization for \( k \)-means is used to generate various base clusterings in each ensemble model. The sigma \( (\sigma) \) is obtained from the grid search scheme to calculate kernel-induced Euclidean distances for the Gaussian kernel function: 0.5*var for ‘Year 4 CE’, 0.6*var for ‘Year 4 CS’, and 0.62*var for Iris, where var is the variance of the corresponding data set. For comparison, we reconsider the methods used in Vo and Nguyen (2018) and add the LWEA ensemble method proposed in Huang et al. (2017). Their brief descriptions are given as follows:

- **k-means** (MacQueen, 1967): this method is a base algorithm used in other ensemble methods.
- **OOA**: this ensemble method is popular with the co-association matrix, where each cell value \((i, j)\) is a cumulative number of co-occurrences in the same cluster of the two corresponding objects \(X_i\) and \(X_j\).
- **LWEA** (Huang et al., 2017): this ensemble method was proposed with a locally weighted co-association matrix instead of the traditional co-association matrix.
- **BOCA** (Pattanodom et al., 2016): this ensemble method was presented with the binary object-cluster association (BOCA) matrix.
- **WOCA** (Vo & Nguyen, 2018): this ensemble method was defined with the weighted object-cluster association (WOCA) matrix instead of the binary one.

For effectiveness measurement, we used the external cluster validation scheme with Normalized Mutual Information (NMI). This measure is widely used and available for consensus clustering as presented in Huang et al. (2017). Its larger value signifies that the resulting clustering is better, i.e. the clustering method is more effective. For more stable experimental results with less randomness, each experiment was run 50 times. Their averaged values were recorded in Tables 2–4 for ‘Year 4 CE’, ‘Year 4 CS’, and Iris data sets, respectively. The best results are presented in bold. In addition, Table 5 is provided for the experimental results with \( \sigma \)'s different values on three data sets when the ensemble size \( e \) is 50 for all the experiments in this group.

With these experimental results, our study focuses on four following questions:

---

Table 2. NMI values on the ‘Year 4 CE’ data set with \( e \)'s varying values.

| \( e \) | 10  | 20  | 30  | 40  | 50  | 60  | 70  | 80  | Average |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|---------|
| \( k \)-means | 0.45 | 0.46 | 0.47 | 0.44 | 0.44 | 0.46 | 0.45 | 0.45 | 0.45 |
| OOA   | 0.41 | 0.40 | 0.40 | 0.40 | 0.41 | 0.41 | 0.40 | 0.40 | 0.40 |
| LWEA  | 0.43 | 0.41 | 0.40 | 0.40 | 0.42 | 0.39 | 0.40 | 0.39 | 0.41 |
| BOCA  | 0.41 | 0.41 | 0.38 | 0.41 | 0.39 | 0.41 | 0.40 | 0.39 | 0.40 |
| WOCA  | 0.50 | 0.49 | 0.50 | 0.49 | 0.50 | 0.49 | 0.50 | 0.50 | 0.50 |
| KWOCA | **0.51** | **0.52** | **0.51** | **0.51** | **0.51** | **0.52** | **0.51** | **0.51** | **0.51** |
How well does the proposed ensemble method create the cluster models compared to its base clustering method?

How well does the proposed ensemble method create the cluster models compared to the other methods using the traditional co-association matrix, the locally weighted co-association matrix, and the binary object-cluster association matrix?

How well is the proposed ensemble method improved with the kernel-induced weighted object-cluster association matrix, compared to its current version, WOCA, which used the weighted object-cluster association matrix?

How stable is the proposed ensemble method with respect to its parameters?

For the first question, kWOCA has much better NMI values than its base clustering method. This reflects the well-known feature of an ensemble model compared to a single one. While the results of k-means are different from execution to execution due to the randomness in its initialization, those of kWOCA are more stable on all the used data sets. Nevertheless, the simplicity and efficiency of k-means are significant for us to select it for generating base clusterings. Consensus clustering can then combine them and finalize a more effective clustering.

### Table 3. NMI values on the ‘Year 4 CS’ data set with $e$’s varying values.

| $e$ | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | Average |
|-----|----|----|----|----|----|----|----|----|---------|
| $k$-means | 0.15 | 0.15 | 0.16 | 0.16 | 0.15 | 0.16 | 0.15 | 0.16 | 0.16 |
| OOA | 0.13 | 0.15 | 0.15 | 0.15 | 0.16 | 0.15 | 0.15 | 0.15 | 0.15 |
| LWEA | 0.12 | 0.16 | 0.16 | 0.1 | 0.16 | 0.14 | 0.16 | 0.14 | 0.14 |
| BOCA | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.13 | 0.15 | 0.14 |
| WOCA | 0.18 | 0.18 | 0.18 | 0.18 | 0.18 | 0.18 | 0.17 | 0.18 | 0.18 |
| kWOCA | **0.24** | **0.25** | **0.25** | **0.25** | **0.25** | **0.24** | **0.24** | **0.24** | **0.25** |

### Table 4. NMI values on the Iris data set with $e$’s varying values.

| $e$ | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | Average |
|-----|----|----|----|----|----|----|----|----|---------|
| $k$-means | 0.72 | 0.72 | 0.69 | 0.71 | 0.71 | 0.72 | 0.71 | 0.72 | 0.71 |
| OOA | 0.71 | 0.71 | 0.71 | 0.72 | 0.71 | 0.70 | 0.72 | 0.71 | 0.71 |
| LWEA | 0.68 | 0.74 | 0.74 | 0.75 | 0.74 | 0.74 | 0.74 | 0.74 | 0.73 |
| BOCA | 0.70 | 0.68 | 0.70 | 0.69 | 0.72 | 0.70 | 0.68 | 0.69 | 0.69 |
| WOCA | 0.76 | 0.77 | 0.76 | 0.76 | 0.76 | 0.77 | 0.77 | 0.77 | 0.77 |
| kWOCA | **0.79** | **0.79** | **0.78** | **0.78** | **0.78** | **0.78** | **0.79** | **0.79** | **0.79** |

### Table 5. The resulting NMI values on three data sets with $\sigma$’s varying values using different ratios and $e$=50.

| No | CE Year 4 | CS Year 4 | Iris |
|----|-----------|-----------|------|
|    | Ratio for $\sigma(r_{c})$ | NMI | Ratio for $\sigma(r_{c})$ | NMI | Ratio for $\sigma(r_{c})$ | NMI |
| 1  | 0.1 | 0.51 | 0.1 | 0.24 | 0.54 | 0.77 |
| 2  | 0.2 | 0.51 | 0.2 | 0.24 | 0.55 | 0.78 |
| 3  | 0.3 | 0.51 | 0.3 | 0.24 | 0.56 | 0.78 |
| 4  | 0.4 | 0.51 | 0.4 | 0.24 | 0.57 | 0.77 |
| 5  | 0.5 | 0.51 | 0.5 | 0.25 | 0.58 | 0.78 |
| 6  | 0.6 | 0.51 | 0.6 | 0.25 | 0.59 | 0.78 |
| 7  | 0.7 | 0.51 | 0.7 | 0.25 | 0.6 | 0.77 |
| 8  | 0.8 | 0.51 | 0.8 | 0.25 | 0.61 | 0.78 |
| 9  | 0.9 | 0.51 | 0.9 | 0.24 | 0.62 | 0.78 |
| 10 | 1   | 0.51 | 1   | 0.25 | 0.63 | 0.79 |
| Average | **0.51** | **0.25** | **0.78** |
To answer the second question, kWOCA is compared to OOA, LWEA (Huang et al., 2017), and BOCA (Pattanodom et al., 2016). Experimental results show that with different ensemble sizes, NMI values from kWOCA are always higher than those from the others on all the used data sets. The difference ranges from about 10% to about 80%. The most difference is with the ‘CS Year 4’ data set and the least one with the Iris data set. This is understandable because the overlapping and sparseness characteristics of each data set have a certain impact on the data that we have extracted in the ensemble space. In particular, the Iris data set is a firm complete one with three very well separated classes while the educational data sets are not. Although these data sets have three known classes, they are sparse and their instances are not well discriminated. Nonetheless, kWOCA can generate better clusters for these instances with better cluster quality improvement.

For the third question, the affect of the data characteristics gets clearer because there is much more difference between kWOCA and WOCA (Vo & Nguyen, 2018) on the ‘CS Year 4’ data set while just a bit one on the ‘CE Year 4’ and Iris data sets. kWOCA has two extra features as improved from WOCA. The first feature is to relax the number of clusters in consensus clustering; instead, determining a value for that parameter with BIC values. The second one is to redefine the object-cluster association weight for each cell in the data matrix in consensus clustering using kernel-induced Euclidean distances and level-wise BIC-based ratios. The experimental results on the educational data sets confirm the effectiveness of these two features while those on the Iris data set confirm the effectiveness of the second one. For the first feature, kWOCA can improve WOCA on the educational data sets when using varying values for the number of clusters in consensus clustering. The differences in NMI values from the ‘CS Year 4’ data set are a lot higher than those from the ‘CE Year 4’ data set because kWOCA returned six to eight clusters for the previous one and two or three clusters for the latter, while WOCA was executed with three clusters in consensus clustering for both data sets. This also confirms the well-known capability of the BIC-based model selection and reflects the overlapping in the ‘CS Year 4’ data set which should be clustered with more than 3 groups. For the second feature, the results on the Iris data set introduce its effectiveness alone with no contribution of the first feature, while those on the educational data sets show its effectiveness in combination with the first feature. When we weight the differences between the clusters in each base clustering and also between the base clusterings, the instances can be more discriminated. Besides, kernel-induced Euclidean distances also help expanding the ensemble space into a higher dimensional space so that the instances can be more separated from each other. As a result, new weights can give more details about each association between an instance and a cluster in a base clustering once consensus clustering is conducted. kWOCA is thus an improved version of WOCA.

As for the fourth question, our experimental results present the stable effectiveness of kWOCA with respect to $e$ and $\sigma$, which are its two important parameters. Shown in Tables 2–4, as we fix $\sigma$, varying $e$ gave the quite similar NMI values for each data set. Shown in Table 5, as we fix $e$, varying $\sigma$ did too. Therefore, a grid search scheme can be used with kWOCA to determine appropriate values for $e$ and $\sigma$ with ease.

Above all, Paired Samples T-Test is conducted to check statistically significant differences in the NMI results from 50 runs of kWOCA and from those of the others on three datasets. Confidence interval percentage was set to 95%. Sig.(2–tailed) values were
recorded in Table 6. In each run on each dataset, an ensemble size is 50, the number of clusters is set to 3 for the Iris dataset with all the methods, and the number of clusters is automatically defined for the educational datasets with kWOCA and 3 with all the other methods. Fifty was chosen for the ensemble size in the experiments because it is large enough for statistical tests. Moreover, in the previous experiments, the NMI values associated with this ensemble size were almost the same as the averaged ones.

In Table 6, the statistically significant difference in NMI results between kWOCA and another method \( X \) is examined. It is denoted as kWOCA > \( X \). The test results confirm the more effectiveness of kWOCA on all three datasets as compared to the other methods with Sig. = .000. Such results showed that kWOCA has been better than its base clustering method and other ensemble methods in almost all the cases. However, it was found that the difference in NMI values between kWOCA and WOCA couldn’t be realized for the ‘CE Year 4’ dataset, while there are statistically significant differences between kWOCA and WOCA for the ‘CS Year 4’ and Iris data sets. This can be explained with the data shortage of the ‘CE Year 4’ dataset, leading to the inapplicability of statistical methods. Indeed, the ‘CE Year 4’ dataset has a limited number of instances in a huge space, particularly 186 instances in a 43-dimension space. Regardless of the results on the ‘CE Year 4’ dataset, the extended features defined for kWOCA have their own merits to advance consensus clustering with richer information in a kernel-induced feature space.

In short, our kWOCA ensemble method with a kernel-induced weighted object-cluster co-association matrix outperforms its base clustering method and the existing ones. It can produce better clusters on a consistent basis. Experimental results also show the stability of the proposed method with respect to its parameters. As a result, kWOCA can be utilized in practice to find the groups of our similar students. Furthermore, cluster analysis can be made to determine the characteristics of each student group. Based on those, appropriate support can be given to each group in time.

### 5. Conclusions

In this paper, a kernel-induced weighted object-cluster association-based ensemble method, kWOCA, has been proposed for clustering undergraduate students based on their study performance. This kWOCA method is a novel solution to the student clustering task that can return non-linear clusters with no requirement about the predefined number of desired clusters. In addition, more discrimination between the objects and between the clusters in a base clustering and between the base clusterings in the ensemble has been defined and encoded into our kernel-induced weighted object-cluster association matrix for consensus clustering from many available base clusterings. Using this more informative and comprehensive association matrix, better ensemble clusterings can be achieved. Experimental results on both real and benchmarking data sets have confirmed that

| Data set     | kWOCA > k-means | kWOCA > OOA | kWOCA > LWEA | kWOCA > BOCA | kWOCA > WOCA |
|--------------|-----------------|-------------|--------------|--------------|--------------|
| CE Year 4    | .000            | .000        | .000         | .000         | .892         |
| CS Year 4    | .000            | .000        | .000         | .000         | .000         |
| Iris         | .000            | .000        | .000         | .000         | .000         |
kWOCA outperforms the other existing ensemble approaches as $k$-means is used as their base clustering algorithm. Our empirical study has reported higher NMI values for the resulting clusterings from kWOCA. Through these better clusters in consensus clustering, our similar students can be linked together and further cluster analysis can be made for their group-based study support.

Although kWOCA has improved our previous WOCA method, our work on these ensemble models is in its infancy. In the future, more consideration is given to sparse data handling to add the robustness into our method. Besides, a parameter-free version of our method will be investigated for its practical use in educational decision support models and systems.

**Acknowledgements**

Many thanks go to the authors of Huang et al. (2017) for the codes.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**Funding**

This research is funded by Vietnam National University Ho Chi Minh City, Vietnam, under grant number C2017–20–18.

**Notes on contributors**

Chau Thi Ngoc Vo is currently with Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology, Vietnam National University-HCMC, Vietnam. She got a bachelor’s degree in Information Technology from Ho Chi Minh City University of Technology, Vietnam National University-HCMC, Vietnam, in 2003, a master’s degree and a doctoral degree from King Mongkut’s Institute of Technology Ladkrabang, Thailand, in 2005 and 2008, respectively, under the AUN/Seed-Net scholarships from JICA. Her research of interest includes complex data modelling and analysis, data mining and knowledge discovery, decision support systems and other modern information systems.

Phung Hua Nguyen is a senior lecturer in the Faculty of Computer Science and Engineering, HCMC University of Technology, Vietnam National University - HCMC, Vietnam. He has received his Ph.D. (2005) in Computer Science from UNSW, Australia, and M.Eng (1999) and B.Eng in Computer Engineering from HCMC University of Technology, VNU-HCMC, Vietnam. His research interests include scheduling, data mining, program analysis and verification.

**References**

Academic Affairs Office. (2017, June 19). Ho Chi Minh City University of Technology, Vietnam. Retrieved from [http://www.aao.hcmut.edu.vn](http://www.aao.hcmut.edu.vn)

Adjei, S., Ostrow, K., Erickson, E., & Heffernan, N. (2017). *Clustering students in ASSISTments: Exploring system and school-level traits to advance personalization*. Proceedings of the 10th international conference on Educational Data Mining (pp. 340–341).
Bresfelean, V. P., Bresfelean, M., & Ghisoiu, N. (2008). Determining students’ academic failure profile founded on data mining methods. Proceedings of the ITI 2008 30th international conference on Information Technology Interfaces (pp. 317–322).

Campagni, R., Merlini, D., & Verri, M. C. (2014). Finding regularities in courses evaluation with k-means clustering. Proceedings of the 6th international conference on Computer Supported Education (pp. 26–33).

Franek, L., & Jiang, X. (2014). Ensemble clustering by means of clustering embedding in vector spaces. Pattern Recognition, 47(2), 833–842.

Huang, D., Wang, C.-D., & Lai, J.-H. (2017). Locally weighted ensemble clustering. IEEE Transactions on Cybernetics, 1–14.

Iris. (2017, June 11). UCI Machine Learning Repository. Retrieved from http://archive.ics.uci.edu/ml

Jayabal, Y., & Ramanathan, C. (2014). Clustering students based on student’s performance – a Partial Least Squares Path Modeling (PLS-PM) study. Proceedings of the MLDM, LNAI 8556 (pp. 393–407).

Jovanovic, M., Vukicevic, M., Milovanovic, M., & Minovic, M. (2012). Using data mining on student behavior and cognitive style data for improving e-learning systems: A case study. International Journal of Computational Intelligence Systems, 5, 597–610.

Kerr, D., & Chung, G. K. W. K. (2012, October). Identifying key features of student performance in educational video games and simulations through cluster analysis. Journal of Educational Data Mining, 4(1), 144–182.

Liu, H., Wu, J., Liu, T., Tao, D., & Fu, Y. (2017). Spectral ensemble clustering via weighted k-means: Theoretical and practical evidence. IEEE Transactions on Knowledge and Data Engineering, 29(5), 1129–1143.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability (pp. 1:281–1:297).

Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning, 52, 91–118.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems, 14, 1–8.

Pattanodom, M., Iam-On, N., & Boongoen, T. (2016). Clustering data with the presence of missing values by ensemble approach. Proceedings of the 2nd Asian conference on Defence Technology (pp. 114–119).

Pelleg, D., & Moore, A. (2000). X-means: Extending K-means with efficient estimation of the number of clusters. Proceedings of the 17th international conference on Machine Learning (pp. 727–734).

Ren, Y., Domeniconi, C., Zhang, G., & Yu, G. (2016). Weighted-object ensemble clustering: methods and analysis. Knowledge and Information Systems, 1–29. doi:10.1007/s10115-016-0988-y

Topchy, A., Jain, A. K., & Punch, W. (2005). Clustering ensembles: Models of consensus and weak partitions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(12), 1866–1881.

Vo, T. N. C., & Nguyen, H. P. (2018). A weighted object-cluster association-based ensemble method for clustering undergraduate students. In N. T. Nguyen, D. H. Hoang, T.-P. Hong, H. Pham, & B. Trawiński (Eds.), Proceedings of the 10th Asian Conference on Intelligent Information and Database Systems (ACID 2018), LNAI 10751 (pp. 587–598).

Wu, J., Liu, H., Xiong, H., Cao, J., & Chen, J. (2015). K-means-based consensus clustering: A unified view. IEEE Transactions on Knowledge and Data Engineering, 27(1), 155–169.

Zhong, C., Yue, X., Zhang, Z., & Lei, J. (2015). A clustering ensemble: Two-level-refined co-association matrix with path-based transformation. Pattern Recognition, 48(8), 2699–2709.

Zhou, Z. H. (2012). Ensemble methods: Foundations and algorithms. Boca Raton, FL: Chapman and Hall/CRC.
Appendix. Consensus matrix-based visualization

In this appendix, consensus matrix-based visualization is supplied according to the graphical presentation introduced in Monti, Tamayo, Mesirov, and Golub (2003). The visualized results give us better understanding of our datasets. They have no reflection on the results of our method because in our method, each consensus matrix has been further enriched with multi-level weights in a kernel-induced feature space. For brevity, the ‘CE Year 4’ dataset is used for visualization.

First, we obtained consensus matrices with a few varying numbers of clusters: $k = 2 \ldots 10$. For each $k$, each consensus matrix is a result of 50 runs of the base clustering method, $k$-means. Each consensus matrix is then visualized with a colour-coded heat map with white for 0 and dark blue for 1. Using each consensus matrix, a histogram is constructed with 101 bins in $[0, 1]$ and a gap of 0.01. This histogram is then used to achieve the empirical cumulative distribution (CDF) over the range $[0, 1]$. In addition to this histogram, another histogram is built with the sorted set of $n(n-1)/2$ entries in the consensus matrix where $n$ is the number of instances in the given dataset. Based on the histogram associated with the consensus matrix, the area $A(k)$ under the CDF is calculated for each $k$. After that, the proportion increase in the CDF area, which is called Delta Area, is examined with varying $k$ values. All the related formulas can be checked with those in Monti et al. (2003).

In the following, we present the heat maps stemming from the consensus matrices with $k = 2, 3, 5,$ and 10 in Figures A1–A4, respectively. Among them, $k = 3$ is the true number of clusters in the given dataset. In Figure A5, a chart of corresponding consensus CDFs is displayed. In Figure A6, a Delta Area plot is shown. In this graphical presentation, we can determine the optimal number of clusters for the given dataset and notice the stability in cluster membership of the dataset.

For our dataset, it is realized that the true number of clusters, which is 3, was found with the Delta Area plot. This finding is confirmed by the clearer heat map associated with the consensus matrix with $k = 3$. At the same time, cluster membership of the dataset is not quite stable via the consensus CDF chart. This is consistent with the NMI results that we obtained in the previous experiments. Therefore, more improvement in consensus clustering has been made with WOCA and kWOCA for better clusters.
Figure A1. Color-coded heat map corresponding to the consensus matrix with \( k = 2 \).
Figure A2. Color-coded heat map corresponding to the consensus matrix with $k = 3$. 
Figure A3. Color-coded heat map corresponding to the consensus matrix with $k = 5$. 
Figure A4. Color-coded heat map corresponding to the consensus matrix with $k = 10$. 
**Figure A5.** Empirical CDFs corresponding to the entries of consensus matrices.

**Figure A6.** Proportion increase *Delta Area* in the area under the CDF.