VNU Journal of Science: Comp. Science & Com. Eng., Vol. 33, No. 2 (2017) 66-75

Educational Data Clustering in a Weighted Feature Space Using Kernel *K*-Means and Transfer Learning Algorithms

Vo Thi Ngoc Chau*, Nguyen Hua Phung

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

Abstract

Educational data clustering on the students' data collected with a program can find several groups of the students sharing the similar characteristics in their behaviors and study performance. For some programs, it is not trivial for us to prepare enough data for the clustering task. Data shortage might then influence the effectiveness of the clustering process and thus, true clusters can not be discovered appropriately. On the other hand, there are other programs that have been well examined with much larger data sets available for the task. Therefore, it is wondered if we can exploit the larger data sets from other source programs to enhance the educational data clustering task on the smaller data sets from the target program. Thanks to transfer learning techniques, a transfer-learning-based clustering method is defined with the kernel *k*-means and spectral feature alignment algorithms in our paper as a solution to the educational data clustering task in such a context. Moreover, our method is optimized within a weighted feature space so that how much contribution of the larger source data sets to the clustering process can be automatically determined. This ability is the novelty of our proposed transfer learning-based clustering solution as compared to those in the existing works. Experimental results on several real data sets have shown that our method consistently outperforms the other methods using many various approaches with both external and internal validations.

Received 16 Nov 2017, Revised 31 Dec 2017; Accepted 31 Dec 2017

Keywords: Educational data clustering, kernel *k*-means, transfer learning, unsupervised domain adaptation, weighted feature space.

1. Introduction

Due to the very significance of education, data mining and knowledge discovery have been investigated much on educational data for a great number of various purposes. Among the mining tasks recently considered, data clustering is quite popular for the ability to find the clusters inherent in an educational data set. Many existing works in [4, 5, 11-13, 19] have examined this task. Among these works, [19] is one of our previous works for the same purpose to generate several groups of the students who have similar study performance while the others have been proposed before with the following different purposes. For example, [4] generated and analyzed the clusters for student's profiles, [5] discovered student groups for the regularities in course evaluation, [11] utilized the student groups to find how the study performance has been related to the medium of study in main subjects, [12] found the student groups with similar cognitive styles and grades in an e-learning system, and [13] derived the student groups with similar actions. Except for

* Corresponding authors. E-mails: chauvtn@hcmut.edu.vn
https://doi.org/10.25073/2588-1086/vnucsce.172

[19], none of the aforementioned works considers lack of educational data in their tasks. In our context, data collected with the target program is not large enough for the task. This leads to a need of a new solution to the educational data clustering task in our context.

Different from the existing works in the educational data clustering research area, our work aims at a clustering solution which can work well on a smaller target data set. In order to accomplish such a goal, our solution exploits another larger data set collected from a source program and then makes the most of transfer learning techniques for a novel method. The resulting method is a Weighted kernel *k*-means (SFA) algorithm, which can discover the clusters in a weighted feature space. This method is based on the kernel *k*-means and spectral feature alignment algorithms with a new learning process including the automatic adjustment of the enhanced feature space once running transfer learning at the representation level on both target and source data sets.

As compared to the existing unsupervised transfer learning techniques in [8, 15] where transfer learning was conducted at the instance level, our method is more appropriate for educational data clustering. As compared to the existing supervised techniques in [14, 20] on multiple educational data sets, their mining tasks were dedicated to classification and regression, respectively, not to clustering. On the other hand, transfer learning in [20] is also different from ours as using Matrix Factorization for sparse data handling.

In comparison with the existing works in [3, 6, 9, 10, 17, 21] on domain adaptation and transfer learning, our method not only applies an existing spectral feature alignment algorithm (SFA) in [17] but also advances the contribution of the source data set to our unsupervised learning process, i.e. our clustering process for the resulting clusters of higher quality. In particular, [6] used a parallel data set to connect the target domain with the source domain instead of using domain-independent features called in [17] or pivot features called in [3, 21]. In practice, it is non-trivial to prepare such a parallel data set in many different application domains, especially those new to transfer learning, like the educational domain. Also, not asking for the optimal dimension of the common subspace, [9] defined the Heterogeneous Feature Augmentation (HFA) method to obtain new augmented feature representations using different projection matrices. Unfortunately, these projection matrices had to be learnt with both labeled target and source data sets while our data sets are unlabeled. Therefore, HFA is not applicable to our task. As for [10], a feature space remapping method is defined to transfer knowledge from domains to domains using meta-features via which the features of the target space can be connected with those of the source one. Nevertheless, [10] then constructed a classifier on the labeled source data set together with the mapped labeled target data set. This classifier would be used to predict instances in the target domain. Such an approach is hard to be considered in our context, where we expect to discover the clusters inherent only in the target space using all the unlabeled data from both target and source domains. In another approach, [21] used joint non-negative matrix factorization to link heterogeneous features with pivot features so that a classifier learnt on a labeled source data set could be used for instances in a target data set. Compared to [21], our work utilizes an unlabeled source data set and does not build a common space where the clusters would be discovered. Instead we construct a weighted feature space for the target domain based on the knowledge transferred from the source domain at the representation level. Different from the aforementioned works, [3, 17] enabled the transfer learning process on unlabeled target and source data at the representation level. Their approaches are very suitable for our unsupervised learning process. While [3] was based on pivot features to generate a common space via structural correspondence learning, [17] was based on domain-independent features to align other domain-specific features from both target and source domains via spectral clustering [16] with Laplacian eigenmaps [2] and spectral graph theory [7]. In [3], many pivot predictors need to be prepared while as a more recent work, [17] is closer to our clustering

task. Nonetheless, [3, 17] required users to pre-specify how much the knowledge can be transferred between two domains via h and K parameters, respectively. Thus, once applying the approach in [17] to unsupervised learning, we decide to change a fixed enhanced feature space with predefined parameters to a weighted feature space which can be automatically learnt along with the resulting clusters.

In short, our proposed method is novel for clustering the instances in a smaller target data set with the help of another larger source data set. The resulting clusters found in a weighted feature space can reveal how the similar students are non-linearly grouped together in their original target data space. These student groups can be further analyzed for more information in support of in-trouble students. The better quality of each student group in the resulting clusters has been confirmed via both internal objective function and external Entropy values on real data sets in our empirical study.

The rest of our paper is organized as follows. Section 2 describes an educational data clustering task of our interest. In section 3, our transfer learning-based kernel k-means method in a weighted feature space is proposed. We then present an empirical study with many experimental results in order to evaluate the proposed method in comparison with the others in section 4. Finally, section 5 concludes this paper and states our future works.

2. An educational data clustering task for grouping the students

Grouping the students into several clusters each of which contains the most similar students is one of the popular educational data mining tasks as previously introduced in section 1. In our paper, we examine this task in a more practical context where a smaller data set can be prepared for the target program. Some reasons for such data shortage can be listed as follows. Data collection got started late for data analysis requirements. Data digitization took time for a larger data set. The target program is a young one with a short history. As a result, data in a data space where our students are modeled is limited, leading to inappropriate clusters discovered in a small set of the target program.

Supporting the task to form the clusters of really similar students in such a context, our work takes advantage of the existing larger data sets from other source program. This approach distinguishes our work from the existing ones in the educational data mining research area for the clustering task. In the following, our task is formally defined in this context.

Let **A** be our target program associated with a smaller data set D_t in a data space characterized by the subjects which the students must accomplish for a degree in program **A**. Let **B** be another source program associated with a larger data set D_s in another data space also characterized by the subjects that the students must accomplish for a degree in program **B**.

In our input, D_t is defined with n_t instances each of which has $(t+p)$ features in the $(t+p)$-dimensional vector space where t features stem from the target data space and p features from the shared data space between the target and source ones.

$$D_t = \{X_r, \forall\, r=1..n_t\} \quad (1)$$

where X_r is a vector: $X_r = (x_{r,1}, .., x_{r,(t+p)})$ with $x_{r,d} \in [0, 10]$, $\forall\, d=1..(t+p)$

In addition, D_s is defined with n_s instances each of which has $(s+p)$ features in the $(s+p)$-dimensional vector space where s features stem from the source data space. It is noted that D_t is a smaller target data set and D_s is a larger source data set in such a way that: $n_t << n_s$.

$$D_s = \{X_r, \forall\, r=1..n_s\} \quad (2)$$

where X_r is a vector: $X_r = (x_{r,1}, .., x_{r,(s+p)})$ with $x_{r,d} \in [0, 10]$, $\forall\, d=1..(s+p)$

As our output, the clusters of the instances in D_t are discovered and returned. It is expected that the resulting clusters are of higher quality once the clustering process is executed on both D_t and D_s as compared to those with the clustering process on only D_t. Each cluster represents a group of the most similar students sharing the similar performance characteristics. Besides, each cluster is quite well separated

from each other so that dissimilar students can be included into different clusters.

Exploiting D_s with transfer learning techniques and kernel *k*-means, our clustering method is defined with a clustering process in a weighted feature space instead of a traditional data space of either D_t or D_s. The weighted feature space is learnt automatically according to the contribution of the source data set. It is expected that this process can do clustering more effectively in the weighted feature space.

3. The proposed educational data clustering method in a weighted feature space

In this section, our proposed educational data clustering method in a weighted feature space is defined using kernel *k*-means [18] and the spectral feature alignment algorithm [17]. It is named "Weighted kernel *k*-means (SFA)". Our method first constructs a feature space from the enhancement of new spectral features derived from the feature alignment between the target and source spaces with respect to their domain-independent features. Using this new feature space, it is non-trivial for us to determine how much the new spectral features contribute to the existing target space for the clustering process. Therefore, our method includes the adjusting of the new feature space towards the best convergence of the clustering process. In such a manner, this new feature space is called a weighted feature space. In this weighted feature space, kernel *k*-means is executed for more robust arbitrarily-shaped clusters as compared to traditional *k*-means.

3.1. A Weighted Feature Space

Let us first define the target data space as S_t and the new weighted feature space as S_w. S_t has $(t+p)$ dimensions where t dimensions corresponds to t domain-specific features of the target data set D_t and p dimensions corresponds to p domain-independent features shared by the target data set D_t and the source data set D_s. In the target data space S_t, every dimension is treated equally to each other. Different from S_t, S_w has $(t+2*p)$ dimensions where $(t+p)$ dimensions are inherited from the target data space S_t and the remaining p dimensions are all the new spectral features obtained from both target and source data spaces using the SFA algorithm. In addition, every feature at the d-th dimension in S_w has a certain degree of importance, reflected by a weight w_d, in representing an instance in the space and then in discriminating an instance from the others in the clustering process. These weights are normalized so that their total sum can be 1. At the instance level, each instance in D_t is mapped to a new instance in S_w using the feature alignment mapping φ learnt with the SFA algorithm. A collection of all the new instances in S_w forms our enhanced instance set D_w which is then used in the learning process to discover the clusters. D_w is formally defined as follows:

$$D_w = \{X_r, \forall\ r=1..n_t\} \quad (3)$$

where X_r is a vector: $X_r = (x_{r,1}, .., x_{r,(t+p)}, \varphi(X_r))$ with $x_{r,d} \in [0, 10]$, $\forall\ d=1..(t+p)$ stemming from the original ones and $\varphi(X_r)$ is a p-dimensional vector for p new spectral features.

The new weighted feature space captures the support transferred from the larger source data set for the clustering process on the smaller target data set. In order to automatically determine the importance of each feature in S_w, the clustering process not only learns the clusters inherent in the target data set D_t via the enhanced set D_w but also optimizes the weights of S_w to better generate the clusters.

3.2. The Clustering Process

Playing an important role, the clustering process shows how our method can discover the clusters in the target data set. Based on kernel *k*-means with a predefined number k of desired clusters, it is carried out with respect to minimizing the value of the following objective function in the weighted feature space S_w:

$$J^{\Phi}(D_w, C^{\Phi}) = \sum_{r=1..n_t} \sum_{o=1..k} \gamma_{or} \parallel \Phi(X_r) - C_o \parallel^2 \quad (4)$$

where γ_{or} shows the membership of X_r with respect to the cluster C_o: 1 if a member and otherwise, 0. C_o is a cluster center in S_w with an implicit mapping function Φ, defined below:

$$C_o = \frac{\sum_{q=1..n_t} \gamma_{oq} \Phi(X_q)}{\sum_{q=1..n_t} \gamma_{oq}} \quad (5)$$

As we never decide the function Φ explicitly, a kernel trick is made the most of. Due to popularity, the Gaussian kernel function is used in our work. It is defined in (6) as follows:

$$K(X_i, X_j) = e^{-\frac{\|X_i - X_j\|^2}{2\sigma^2}} \quad (6)$$

where X_i and X_j are two vectors and σ is a bandwidth of the kernel function.

With the Gaussian kernel function, a kernel matrix *KM* is computed on the enhanced data set D_w in the weighted feature space S_w as follows:

$$KM(X_r, X_q) = K_{rq} = e^{-\frac{\|X_r - X_q\|^2}{2\sigma^2}}$$

$$KM(X_r, X_q) = K_{rq} = e^{-\frac{\sum_{d=1..t+2*p} w_d^2 (x_{r,d} - x_{q,d})^2}{2\sigma^2}} \quad (7)$$

for r=1..n_t and q=1..n_t.

In our clustering process, a weight vector (w_1, w_2, …, w_d, …, w_{t+2*p}) for d=1..t+2*p needs to be estimated, leading to the estimation of the kernel matrix *KM* iteratively.

Using the kernel matrix, the corresponding objective function derived from (4) is now shown in the formula (8) as follows:

$$J^{\Phi}(D_W, C^{\Phi}) = \sum_{r=1..n_t} \sum_{o=1..k} \gamma_{or} \left(K_{rr} - \frac{2 \sum_{q=1..n_t} \gamma_{oq} K_{rq}}{\sum_{q=1..n_t} \gamma_{oq}} + \frac{\sum_{v=1..n_t} \sum_{z=1..n_t} \gamma_{ov} \gamma_{oz} K_{vz}}{\sum_{v=1..n_t} \sum_{z=1..n_t} \gamma_{ov} \gamma_{oz}} \right) \quad (8)$$

where we have got K_{rr}, K_{rq}, and K_{vz} in the kernel matrix. γ_{or}, γ_{oq}, γ_{ov}, and γ_{oz} are memberships of the instances X_r, X_q, X_v, and X_z with respect to the cluster C_o as follows:

$$\gamma_{oq} = \begin{cases} 1, & \text{if } X_q \text{ is a member of } C_o \\ 0, & \text{otherwise} \end{cases}$$

$$\gamma_{ov} = \begin{cases} 1, & \text{if } X_v \text{ is a member of } C_o \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

$$\gamma_{oz} = \begin{cases} 1, & \text{if } X_z \text{ is a member of } C_o \\ 0, & \text{otherwise} \end{cases}$$

The clustering process is iteratively executed in the alternating optimization scheme to minimize the objective function. After an initialization, it first updates the clusters and their members, and then estimates the weight vector using gradient descent. Its steps are sequentially performed as follows:

(1). Initialization

(1.1). Make a random initialization and normalization for the weight vector *w*

(1.2). *k* cluster centers are initialized as the result of the traditional *k*-means algorithm in the initial weighted feature space.

(2). Repeat the following substeps until the terminating conditions are true:

(2.1). Compute the kernel matrix using (7)

(2.2). Update the distance between each cluster center C_o and each instance X_r in the feature space for o=1..k and r=1..n_t

$$\| \Phi(X_r) - C_o \|^2 = K_{rr} - 2 \frac{\sum_{q=1..n_t} \gamma_{oq} K_{rq}}{\sum_{q=1..n_t} \gamma_{oq}} + \frac{\sum_{v=1..n_t} \sum_{z=1..n_t} \gamma_{ov} \gamma_{oz} K_{vz}}{\sum_{v=1..n_t} \sum_{z=1..n_t} \gamma_{ov} \gamma_{oz}} \quad (10)$$

(2.3). Update the membership γ_{oq} between the instance X_r and the cluster center C_o for r=1..n_t and o=1..k

$$\gamma_{oq} = \begin{cases} 1, & \text{if } \| \Phi(X_r) - C_o \|^2 = argmin_{o'=1..k} (\| \Phi(X_r) - C_{o'} \|^2) \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

(2.4). Update the weight vector *w* using the following formulas (12), (13), and (14)

$$w_d = w_d - \eta \frac{\partial J^{\Phi}(D_W, C^{\Phi})}{\partial w_d} \quad (12)$$

where d=1..t+2*p and η is a learning rate to control the speed of the learning process.

From (7), we obtain the partial derivative of K_{rq} with respect to w_d for d = 1..t+2*p in the formula (13) as follows:

$$\frac{\partial K_{rq}}{\partial w_d} = \frac{-w_d (x_{r,d} - x_{q,d})^2}{\sigma^2} K_{rq} \quad (13)$$

Using (13), we obtain the partial derivative of $J^{\Phi}(D_w, C^{\Phi})$ with respect to w_d for d = 1..t+2*p in the following formula (14):

$$\frac{\partial J^{\Phi}(D_{\Phi}, C^{\Phi})}{\partial w_d} = \sum_{r=1..n_t} \sum_{o=1..k} \frac{w_d \gamma_{or}}{\sigma^2} \left(2 \frac{\sum_{q=1..n_t} \gamma_{oq} K_{rq} (x_{r,d} - x_{q,d})^2}{\sum_{q=1..n_t} \gamma_{oq}} - \frac{\sum_{v=1..n_t} \sum_{z=1..n_t} \gamma_{ov} \gamma_{oz} K_{vz} (x_{v,d} - x_{z,d})^2}{\sum_{v=1..n_t} \sum_{z=1..n_t} \gamma_{ov} \gamma_{oz}} \right) \quad (14)$$

(2.5). Perform the normalization of the weight vector w in [0, 1]

Once bringing this learning process to our educational domain, we simplify the process so that our method can require only one parameter k which is popularly known for k-means-based algorithms. For other domains, grid search can be used to appropriately choose the following other parameter values. In particular, the bandwidth σ of the kernel function is derived from the variance of the target data set. In addition, the learning rate η is defined as a decreasing function of time instead of a constant specified by users:

$$\eta = \frac{0.01}{1 + iteration\#} \quad (15)$$

where *iteration#* is the current number of iterations.

Regarding the convergence of this process in connection with its terminating conditions, the stability of the clusters discovered so far is used. Due to the nature of the alternating optimization scheme, our learning process sometimes reaches local convergence.

Nonetheless, it can find the clusters in the weighted feature space more effectively as compared to its base clustering process. Indeed, the resulting clusters are better formed in arbitrary shapes in the target data space. They are also more compact and better separated from each other, i.e. of higher quality.

3.3. Characteristics of the Proposed Method

First of all, we would like to make a clear distinction between this work and our previous one in [19]. They have taken into account the same task in the same context using the same base techniques: kernel k-means and the spectral feature alignment algorithm. Nevertheless, this work addresses the contribution of the source data set to the learning process on the target data set at the representation level via a weighted feature space. The weighted feature space is also learnt within the learning process towards the minimization of the objective function of the kernel k-means algorithm. This solution is novel for the task and also makes its initial version in [19] more practical to users.

As including the adjustment of the weighted feature space into the learning process, our current method has more computational cost than the one in [19]. More space is needed for the weight vector w and more computation for updating the kernel matrix KM and the weight vector in each iteration in a larger feature space S_w as compared to those in [19].

In comparison with the other existing works on educational data clustering, our work along with [19] is one of the first works bringing kernel k-means to discover better true clusters of the students which are non-linearly separated. This is because most of the works on educational data clustering such as [4, 5, 12] were based on k-means. In addition, we have addressed the data insufficiency in the task with transfer learning while the others [4, 5, 11-13] did not or [14, 20] exploited multiple data sources for educational data classification and regression tasks in different approaches.

Like [19], this work has defined a transfer learning-based clustering approach different

from those in [8, 15]. In [8], self-taught clustering was proposed and is now a popular unsupervised transfer learning algorithm. The main difference between our works and [8] is the exploiting of the source data set at different levels of abstraction: [8] at the instance level while ours at the representation level. Such a difference leads to the space where the clusters could be formed: [8] in the data (sub)space with co-clustering while ours in the feature space with kernel *k*-means. Moreover, how much contribution of the source data set is automatically determined in our current work while this issue was not examined in [8]. More recently proposed in [15], another unsupervised transfer learning algorithm has been defined for short text clustering. This algorithm is also considered at the instance level as executed on both target and source data sets and then filtering the instances from the source data set to conclude the final clusters in the target data set. For both algorithms in [8, 15], it was assumed that the same data space was used in both source and target domains. In contrast, our works never require such an assumption.

It is believed that our proposed method has its own merits of discovering the inherent clusters of the similar students based on study performance. It can be regarded as a novel solution to the educational data clustering task.

4. Empirical evaluation

In the previous subsection 3.3, we have discussed the proposed method from the theoretical perspectives. In this section, more discussions from the empirical perspectives are provided for an evaluation of our method.

4.1. Data and experiment settings

Data used in our experiments stem from the student information of the students at Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology, Vietnam, [1] where the academic credit system is running. There are two educational programs in context establishment of the task: Computer Engineering and Computer Science. Computer Engineering is our target program and Computer Science our source program. Each program has 43 subjects that the students have to successfully accomplish for their graduation. A smaller target data set with the Computer Engineering program has 186 instances and a larger source data set with the Computer Science program has 1317 instances. These two programs are close to each other with 32 subjects in common in our work. Three true natural groups of the similar students based on study performance are: studying, graduating, and study-stop. These groups are monitored along the study path of the students from year 2 to year 4 corresponding to the "Year 2", "Year 3", and "Year 4" data sets for each program. Their related details are given in Table 1.

Table 1. Details of the programs

Program	Student#	Subject#	Group#
Computer Engineering (Target, A)	186	43	3
Computer Science (Source, B)	1,317	43	3

For choosing parameter values in our method, we set the number *k* of desired clusters to 3, *sigmas* for the spectral feature alignment and kernel *k*-means algorithms to 0.3**variance* where *variance* is the total sum of the variance for each attribute in the target data. The learning rate is set according to (15). For parameters in the methods in comparison, default settings in their works are used.

For comparison with our Weighted kernel *k*-means (SFA) method, we have taken into consideration the following methods:

- *k*-means (CS): the traditional *k*-means algorithm executed in the common space (CS) of both target and source data sets
- Kernel *k*-means (CS): the traditional kernel *k*-means algorithm executed in the common space of both data sets
- Self-taught Clustering (CS): the self-taught clustering algorithm in [8] executed in the common space of both data sets
- Unsupervised TL with *k*-means (CS): the unsupervised transfer learning algorithm in [15] executed with *k*-means as the base algorithm in the common space
- *k*-means (SFA): the traditional *k*-means algorithm executed on the target data set

enhanced with all the 32 new features from the SFA algorithm with no weighting

- Kernel *k*-means (SFA): the traditional kernel *k*-means algorithm executed on the target data set enhanced with all the 32 new features from SFA with no weighting

In order to avoid randomness in execution, 50 different runs of each experiment were prepared and the same initial values were used for all the algorithms in the same experiment. Each experimental result recorded in the following tables is an averaged value. For simplicity, their corresponding standard deviations are excluded from the paper.

For cluster validation in comparison, the averaged objective function and Entropy measures are used. The averaged objective function value is the conventional one in the target data space averaged by the number of attributes. The Entropy value is the total sum of the Entropy value of each resulting cluster in a clustering, calculated according to the formulae in [8]. The averaged objective function measure is an internal one while the Entropy measure is an external one. Both measures are with the smaller values for the better clusters.

4.2. Experimental Results and Discussions

In the following tables Table 2-4, the experimental results corresponding to the data sets "Year 2", "Year 3", and "Year 4" are presented. The best ones are displayed in bold.

Table 2. Results on the "Year 2" data set

Method	Objective Function	Entropy
k-means (CS)	613.83	1.22
Kernel *k*-means (CS)	564.94	1.10
Self-taught Clustering (CS)	553.64	1.27
Unsupervised TL with *k*-means (CS)	542.04	1.01
k-means (SFA)	361.80	1.12
Kernel *k*-means (SFA)	323.26	0.98
Weighted kernel *k*-means (SFA)	**309.25**	**0.96**

Table 3. Results on the "Year 3" data set

Method	Objective Function	Entropy
k-means (CS)	673.60	1.11
Kernel *k*-means (CS)	594.56	0.93
Self-taught Clustering (CS)	923.02	1.46
Unsupervised TL with *k*-means (CS)	608.87	1.05
k-means (SFA)	419.02	0.99
Kernel *k*-means (SFA)	369.37	0.82
Weighted kernel *k*-means (SFA)	**348.44**	**0.78**

Table 4. Results on the "Year 4" data set

Method	Objective Function	Entropy
k-means (CS)	726.36	1.05
Kernel *k*-means (CS)	650.38	0.95
Self-taught Clustering (CS)	598.98	1.03
Unsupervised TL with *k*-means (CS)	555.66	0.81
k-means (SFA)	568.93	0.95
Kernel *k*-means (SFA)	475.57	0.81
Weighted kernel *k*-means (SFA)	**441.71**	**0.74**

Firstly, we check if our clusters can be discovered better in an enhanced feature space using the SFA algorithm than in a common space. In all the tables, it is realized that *k*-means (SFA) outperforms *k*-means (CS) and kernel *k*-means (SFA) also outperforms kernel *k*-means (CS). The differences occur clearly at both measures and show that the learning process has performed better in the enhanced feature space instead of the common space.

This is understandable as the enhanced feature space contains more informative details and thus, a transfer learning technique is valuable for the data clustering task on small target data sets like those in the educational domain.

Secondly, we check if our transfer learning approach using the SFA algorithm is better than other transfer learning approaches in [8, 15]. Experimental results on all the data sets show that our approach with three methods such as *k*-means (SFA), kernel *k*-means (SFA), and Weighted kernel *k*-means (SFA) can help generating better clusters on the "Year 2" and "Year 3" data sets as compared to both approaches in [8, 15]. On the "Year 4" data set, our approach is just better than Self-taught clustering (CS) in [8] while comparable to Unsupervised TL with *k*-means (CS) in [15]. This is because the "Year 4" data set is much denser and thus, the enhancement is just a bit effective. By contrast, the "Year 2" and "Year 3" data sets are sparser with more data insufficiency and thus, the enhancement is more effective. Nevertheless, our method is always better than the others with the smallest values. This fact notes how appropriately and effectively our method has been designed.

Thirdly, we would like to highlight the weighted feature space in our method as compared to both common and traditionally fixed enhanced spaces. In all the cases, our method can discover the clusters in a weighted feature space better than the other methods in other spaces. A weighted feature space can be adjusted along with the learning process and thus help the learning process examine the discrimination of the instances in the space better. It is reasonable as each feature from either original space or enhanced space is important to the extent that the learning process can include it in computing the distances between the instances. The importance of each feature is denoted by means of a weight learnt in our learning process. This property allows forming the better clusters in arbitrary shapes in a weighted feature space rather than a common or a traditionally fixed enhanced feature space.

In short, our proposed method, Weighted kernel *k*-means (SFA), can produce the smallest values for both objective function and Entropy measures. These values have presented the better clusters with more compactness and non-linear separation. Hence, the groups of the most similar students behind these clusters can be derived for supporting academic affairs.

5. Conclusion

In this paper, a transfer learning-based kernel *k*-means method, named Weighted kernel *k*-means (SFA), is proposed to discover the clusters of the similar students via their study performance in a weighted feature space. This method is a novel solution to an educational data clustering task which is addressed in such a context that there is a data shortage with the target program while there exist more data with other source programs. Our method has thus exploited the source data sets at the representation level to learn a weighted feature space where the clusters can be discovered more effectively. The weighted feature space is automatically formed as part of the clustering process of our method, reflecting the extent of the contribution of the source data sets to the clustering process on the target one. Analyzed from the theoretical perspectives, our method is promising for finding better clusters.

Evaluated from the empirical perspectives, our method outperforms the others with different approaches on three real educational data sets along the study path of regular students. Better smaller values for the objective function and Entropy measures have been recorded for our method. Those experimental results have shown the more effectiveness of our method in comparison with those of the other methods on a consistent basis.

Making our method parameter-free by automatically deriving the number of desired clusters inherent in a data set is planned as a future work. Furthermore, we will make use of the resulting clusters in an educational decision support model based on case based reasoning. This combination can provide a more practical but effective decision support model for our educational decision support system. Besides, more analysis on the groups of the students with similar study performance will be done to

create study profiles of our students over the time so that the study trends of our students can be monitored towards their graduation.

Acknowledgements

This research is funded by Vietnam National University Ho Chi Minh City, Vietnam, under grant number C2016-20-16. Many sincere thanks also go to Mr. Nguyen Duy Hoang, M.Eng., for his support of the transfer learning algorithms in Matlab.

References

[1] AAO, Academic Affairs Office, www.aao.hcmut.edu.vn, accessed on 01/05/2017.

[2] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," Neural Computation, vol. 15, no. 6, pp. 1373-1396, 2003.

[3] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," Proc. The 2006 Conf. on Empirical Methods in Natural Language Processing, pp. 120-128, 2006.

[4] V. P. Bresfelean, M. Bresfelean, and N. Ghisoiu, "Determining students' academic failure profile founded on data mining methods," Proc. The ITI 2008 30th Int. Conf. on Information Technology Interfaces, pp. 317-322, 2008.

[5] R. Campagni, D. Merlini, and M. C. Verri, "Finding regularities in courses evaluation with k-means clustering," Proc. The 6th Int. Conf. on Computer Supported Education, pp. 26-33, 2014.

[6] W-C. Chang, Y. Wu, H. Liu, and Y. Yang, "Cross-domain kernel induction for transfer learning," AAAI, pp. 1-7, 2017.

[7] F.R.K. Chung, "Spectral graph theory," CBMS Regional Conf. Series in Mathematics, No. 92, American Mathematical Society, 1997.

[8] W. Dai, Q. Yang, G-R. Xue, and Y. Yu, "Self-taught clustering," Proc. The 25th Int. Conf. on Machine Learning, pp. 1-8, 2008.

[9] L. Duan, D. Xu, and I. W. Tsang, "Learning with augmented features for heterogeneous domain adaptation," Proc. The 29th Int. Conf. on Machine Learning, pp. 1-8, 2012.

[10] K. D. Feuz and D. J. Cook, "Transfer learning across feature-rich heterogeneous feature spaces via feature-space remapping (FSR)," ACM Trans. Intell. Syst. Technol., vol. 6, pp. 1-27, March 2015.

[11] Y. Jayabal and C. Ramanathan, "Clustering students based on student's performance – a Partial Least Squares Path Modeling (PLS-PM) study," Proc. MLDM, LNAI 8556, pp. 393-407, 2014.

[12] M. Jovanovic, M. Vukicevic, M. Milovanovic, M. Minovic, "Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study," Int. Journal of Computational Intelligence Systems, vol. 5, pp. 597-610, 2012.

[13] D. Kerr and G. K.W.K. Chung, "Identifying key features of student performance in educational video games and simulations through cluster analysis," Journal of Educational Data Mining, vol. 4, no. 1, pp. 144-182, Oct. 2012.

[14] I. Koprinska, J. Stretton, and K. Yacef, "Predicting student performance from multiple data sources," Proc. AIED, pp. 1-4, 2015.

[15] T. Martín-Wanton, J. Gonzalo, and E. Amigó, "An unsupervised transfer learning approach to discover topics for online reputation management," Proc. CIKM, pp. 1565-1568, 2013.

[16] A.Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," Advances in Neural Information Processing Systems, vol. 14, pp. 1-8, 2002.

[17] S. J. Pan, X. Ni, J-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," Proc. WWW 2010, pp. 1-10, 2010.

[18] G. Tzortzis and A. Likas, "The global kernel k-means clustering algorithm," Proc. The 2008 Int. Joint Conf. on Neural Networks, pp. 1978-1985, 2008.

[19] C. T.N. Vo and P. H. Nguyen, "A two-phase educational data clustering method based on transfer learning and kernel k-means," Journal of Science and Technology on Information and Communications, pp. 1-14, 2017. (accepted)

[20] L. Voβ, C. Schatten, C. Mazziotti, and L. Schmidt-Thieme, "A transfer learning approach for applying matrix factorization to small ITS datasets," Proc. The 8th Int. Conf. on Educational Data Mining, pp. 372-375, 2015.

[21] G. Zhou, T. He, W. Wu, and X. T. Hu, "Linking heterogeneous input features with pivots for domain adaptation," Proc. The 24th Int. Joint Conf. on Artificial Intelligence, pp. 1419-1425, 2015.