Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

IEEE Trans Cybern. Author manuscript; available in PMC 2017 January 01.

Published in final edited form as:

IEEE Trans Cybern. 2016 January ; 46(1): 181–193. doi:10.1109/TCYB.2015.2399351.

Cluster Prototypes and Fuzzy Memberships Jointly Leveraged Cross-Domain Maximum Entropy Clustering

Pengjiang Qian [Member, IEEE],
School of Digital Media, Jiangnan University, Wuxi 214122, China

Yizhang Jiang [Member, IEEE],
School of Digital Media, Jiangnan University, Wuxi 214122, China

Zhaohong Deng [Senior Member, IEEE],
School of Digital Media, Jiangnan University, Wuxi 214122, China

Lingzhi Hu,
Philips Electronics North America, Cleveland, OH 44143 USA

Shouwei Sun,
School of Digital Media, Jiangnan University, Wuxi 214122, China

Shitong Wang, and
School of Digital Media, Jiangnan University, Wuxi 214122, China

Raymond F. Muzic Jr. [Senior Member, IEEE]
Department of Radiology and Case Center for Imaging Research, University Hospitals, Case Western Reserve University, Cleveland, OH 44106 USA

Pengjiang Qian: qpengjiang@gmail.com

Abstract

The classical maximum entropy clustering (MEC) algorithm usually cannot achieve satisfactory results in the situations where the data is insufficient, incomplete, or distorted. To address this problem, inspired by transfer learning, the specific cluster prototypes and fuzzy memberships jointly leveraged (CPM-JL) framework for cross-domain MEC (CDMEC) is firstly devised in this paper, and then the corresponding algorithm referred to as CPM-JL-CDMEC and the dedicated validity index named fuzzy memberships-based cross-domain difference measurement (FM-CDDM) are concurrently proposed. In general, the contributions of this paper are fourfold: 1) benefiting from the delicate CPM-JL framework, CPM-JL-CDMEC features high-clustering effectiveness and robustness even in some complex data situations; 2) the reliability of FM-CDDM has been demonstrated to be close to well-established external criteria, e.g., normalized mutual information and rand index, and it does not require additional label information. Hence, using FM-CDDM as a dedicated validity index significantly enhances the applicability of CPM-JL-CDMEC under realistic scenarios; 3) the performance of CPM-JL-CDMEC is generally better than, at least equal to, that of MEC because CPM-JL-CDMEC can degenerate into the standard MEC algorithm after adopting the proper parameters, and which avoids the issue of negative transfer; and 4) in order to maximize privacy protection, CPM-JL-CDMEC employs the known cluster prototypes

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and their associated fuzzy memberships rather than the raw data in the source domain as prior knowledge. The experimental studies thoroughly evaluated and demonstrated these advantages on both synthetic and real-life transfer datasets.

Index Terms

Cross-domain clustering; maximum entropy clustering (MEC); privacy protection; transfer learning; validity index

I. Introduction

Clustering analysis is one of the primary branches in pattern recognition. The conventional clustering approaches can be divided into the following major categories: the partition-based [1]–[12], the hierarchy-based [13], [14], the density-based [15]–[18], the grid-based [19]–[21], and the graph-based [22]–[25], etc. Among them, the partition-based approach, such as the well-known K-means [1] and fuzzy C-means (FCM) [2]–[5], is the most popular approach because of its general applicability to real-life problems. Recently, there is a growing interest in another partition-based method, i.e., maximum entropy clustering (MEC) [6], [7], [26], [27], which integrates two theories of probability-based soft partition and information entropy. Karayiannis [6] proposed MEC based on this understanding that the information entropy (such as Shannon's entropy [28], Renyi's entropy [28], and Havrda–Charvat's entropy [28]) is an essential measure of uncertainty or disorder of a system. Consequently, he devised the objective function by combining two items, i.e., one measures the distortion between the samples and the cluster prototypes, and the other is the Shannon's entropy in terms of fuzzy memberships.

MEC aims to search for the optimal partitions via simultaneously minimizing the distortion and maximizing the entropy. Because of the simplicity of its essence as well as the meaningful physical connotation, the research with respect to MEC has triggered extensive interests. We briefly summarized the representatives as follows. Two convergence analyses regarding MEC were studied in [26] and [27]. Quite a few of model improvements related to MEC have also been presented to date. For example, Li and Mukaidono [29] designed a complete Gaussian membership function for MEC; Ghorbani [30] devised a L1-norm-based MEC objective expression with better robustness; Lao *et al.* [31] manipulated the conventional MEC framework into a weighted modality; Yu [32] presented a general fuzzy clustering theory via summarizing several soft-partition clustering prototypes, e.g., MEC and FCM; Yu and Yang [33] further proposed the optimality test and the complexity analysis methods, based on [32], for guaranteeing desirable clustering performance; Wang *et al.* [34] incorporated the concepts of Vapnik's ε-insensitive loss function as well as weight factor into the MEC framework to improve the identification of outliers; and Zhi *et al.* [35] presented a meaningful joint framework by combing fuzzy linear discriminant analysis and the original MEC objective function. In addition, some application studies with respect to MEC were also developed, such as image compression [36], image segmentation [37], and real-time target tracking [38].

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Traditional clustering approaches, such as MEC, K-means, and FCM, usually work well in the ideal condition where the data is sufficient and pure. However, in reality, the noise and interference information is omnipresent in real life. In addition, there are more challenges regarding the data completeness, in light of the fast developments as well as the rapid requirement changes in modern information systems. In particular, in the early course of a new system, it is extremely difficult to collect sufficient reliable data. Such data shortage severely restricts the practicability of clustering algorithm to a large extent.

Several advanced clustering models have been developed to overcome the challenges of data incompleteness and impureness, such as semi-supervised clustering [39]–[42], co-clustering [43]–[45], multitask learning [46]–[48], and transfer learning [49]–[53]. In our view, transfer learning is the most promising one because of its specific mechanism. Transfer learning works in at least two connected data domains, i.e., one source domain and one target domain, and it allows more than one source domain as needed. It first identifies helpful information from the source domain, in the form of either data or knowledge, and then it migrates this information into the target domain to guide the learning procedure. This auxiliary guidance usually enhances the learning performance in the target domain. In the case where current data is insufficient or impure but there is plenty of useful information coming from related fields or previous studies, transfer learning is an appropriate approach that can significantly improves the clustering performance. Up to now, many methodologies regarding transfer learning have also been reported. For instance, Pan and Yang [49] offered us an outstanding survey on transfer learning. References [50]–[54] investigated the transfer learning-based classification methods. The classification problem could be the most extensive research field on transfer learning by now. References [55]–[58] proposed several transfer regression models. Wang *et al.* [59] and Pan *et al.* [60] presented two dimensionality reduction approaches via transfer learning. In addition, References [45], [61]–[63] connected transfer learning with clustering problems and presented several transfer clustering models.

Based on transfer learning, we comprehensively study the cluster prototypes and fuzzy memberships jointly leveraged (CPM-JL) framework for the cross-domain MEC (CDMEC) issue in this paper. The corresponding algorithm is referred to as CPM-JL-CDMEC. Furthermore, we devise the dedicated validity index, i.e., fuzzy membership-based cross-domain difference measurement (FM-CDDM), in order to aid self-adaptive parameter setting in CPM-JL-CDMEC. Owing to this delicate jointly leveraged transfer mechanism as well as the dedicated FM-CDDM validity index, the effectiveness and practicability of CPM-JL-CDMEC are distinctly improved.

The rest of this paper is organized as follows. In Section II, the classical MEC algorithm and transfer learning are reviewed in brief. In Section III, the CPM-JL framework, the CPM-JL-CDMEC algorithm, the convergence analyses, the parameter setting, and the FM-CDDM index are introduced step-by-step. In Section IV, the experimental validations of the correlated algorithms are presented and discussed. The conclusion is summarized in Section V.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

II. Related Work

A. MEC

In a broad sense, MEC implies a series of clustering methods whose objective function is composed of one modality of maximizing entropy. The detailed clustering framework may vary to a certain extent in different algorithms [5], [28]–[34], and herein, we review [5], which is a good representative for this category of approach.

Let $X = \{\boldsymbol{x}_i | \boldsymbol{x}_i \in R^d, i = 1, 2, \ldots, N\}$ denote a given data set, where d is the data dimensionality and N is the data size. Suppose it can be separated into $C (2 \leq C < N)$ clusters according to a similarity measure. The objective function of the classical MEC algorithm can be rewritten as

$$\begin{aligned} &\min_{\boldsymbol{U},\boldsymbol{V}} \left(\sum_{i=1}^{C}\sum_{j=1}^{N} \mu_{ij} \|\boldsymbol{x}_j - \boldsymbol{v}_i\|^2 + \gamma \sum_{i=1}^{C}\sum_{j=1}^{N} \mu_{ij} \ln \mu_{ij} \right) \\ &\text{s. t.} \quad 0 \leq \mu_{ij} \leq 1 \text{ and } \sum_{i=1}^{C} \mu_{ij} = 1 \quad 1 \leq i \leq C, 1 \leq j \leq N \end{aligned} \tag{1}$$

where, $\|\boldsymbol{x}_j - \boldsymbol{v}_i\|^2$ denotes the distance between the pattern $\boldsymbol{x}_j$ and the cluster prototype $\boldsymbol{v}_i$; $\boldsymbol{U} \in R^{C \times N}$ is the membership matrix consisting of $\mu_{ij} (i = 1, \ldots, C; j = 1, \ldots, N)$, and μ_{ij} denotes the membership of the pattern $\boldsymbol{x}_j$ to the cluster prototype $\boldsymbol{v}_i$; $\boldsymbol{V} \in R^{d \times C}$ is the cluster prototype matrix composed of all cluster prototypes $\boldsymbol{v}_i (i = 1, \ldots, C)$; and $\gamma > 0$ is the regularization parameter of the Shannon's entropy.

Using the Lagrange optimization, the update equations of the cluster prototype $\boldsymbol{v}_i$ and the membership μ_{ij} in (1) can be derived as

$$\boldsymbol{v}_i = \frac{\sum_{j=1}^{N} \mu_{ij} \boldsymbol{x}_j}{\sum_{j=1}^{N} \mu_{ij}}, \; i = 1, 2, \ldots, C \tag{2}$$

$$\mu_{ij} = \frac{\exp\left(-\frac{\|\boldsymbol{x}_j - \boldsymbol{v}_i\|^2}{\gamma}\right)}{\sum_{k=1}^{C} \exp\left(-\frac{\|\boldsymbol{x}_j - \boldsymbol{v}_k\|^2}{\gamma}\right)}, \; i = 1, 2, \ldots, C; j = 1, 2, \ldots, N. \tag{3}$$

B. Transfer Learning

Transfer learning aims to improve the learning performance in the target domain by using the reference information from the source domain [49]. The overall framework of transfer learning is illustrated in Fig. 1. In general, there are two kinds of information that the target domain can obtain from the source domain: raw data or knowledge. Raw data in the source domain is the least sophisticated type of prior information. It may be the most common form to sample the source domain dataset and acquire some representatives and their labels. Knowledge in the source domain is another type of advanced information. Because the original data is not always accessible in the source domain, we may need to conclude

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

knowledge from it sometimes. For example, for the purpose of privacy protection, some raw data maybe cannot be opened and other reasons also could cause the raw data cannot be directly employed even they can be opened. For instance, if there are some underlying drifts between two domains and an unexpected negative influence may occur in the target domain if some improper data are adopted from the source domain. This is the so-called phenomenon of negative transfer. To avoid this potential risk, it is a good choice to draw useful knowledge from the source domain instead of the raw data, e.g., the cluster prototypes in the source domain can be regarded as the good references for the target domain.

So far, transfer learning has been successfully applied in various real-world modeling and learning tasks, such as classification, clustering, and regression [50]–[63].

III. CDMEC Based on Transfer Learning

A. Meaningful Transfer Optimization Formulations for CDMEC

As the two most important elements affecting the clustering performance, cluster prototypes and fuzzy memberships are generally rich in information for reference. Therefore, they are reliable knowledge that can be extracted from the source domain for transfer learning. Based on this idea, for the CDMEC issue, we first devise several transfer optimization formulations from the simple to the complex, and then, we eventually present the complete optimization framework.

1) Cluster Prototypes Leveraged Transfer Optimization Formulation

Definition 1: Let $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_C] \in R^{d\times C}$ denote the cluster prototype matrix in the target domain, where $\boldsymbol{v}_i (i = 1, \ldots, C)$ is the estimated cluster prototype, and d and C are the data dimensionality and the cluster number, respectively. Meanwhile, let $\tilde{\boldsymbol{V}} = [\tilde{\boldsymbol{v}}_1, \ldots, \tilde{\boldsymbol{v}}_C] \in R^{d\times C}$ denote the cluster prototype matrix consisting of each known cluster prototype $\tilde{\boldsymbol{v}}_i (i = 1, \ldots, C)$ in the source domain. Furthermore, make λ be a regularization parameter. Then, the cluster prototype leveraged transfer optimization formulation for CDMEC is defined as

$$\Phi(\boldsymbol{V}) = \lambda \sum_{i=1}^{C} \|\boldsymbol{v}_i - \tilde{\boldsymbol{v}}_i\|^2. \quad (4)$$

Equation (4) represents the total gap between the estimated cluster prototypes in the target domain and the known ones in the source domain. The regularization coefficient λ controls the reference values of the cluster prototypes in the source domain to those in the target domain. Usually, the larger the coefficient λ, the greater the reference values are.

2) CPM-JL Transfer Optimization Formulation

Definition 2: Let $\tilde{\boldsymbol{U}} \in R^{C\times N}$ denote the membership matrix composed of $\tilde{\mu}_{ij} (i = 1, \ldots, C; j = 1, \ldots, N)$, where $\tilde{\mu}_{ij}$ denotes the membership of the pattern $\boldsymbol{x}_j$ in the target domain to the known cluster prototype $\tilde{\boldsymbol{v}}_i (i = 1, \ldots, C)$ in the source domain and it can be derived from (3). By introducing a trade-off factor $\eta \in [0, 1]$ and with the other denotations being the same as those in (4), then CPM-JL transfer optimization formulation can be defined as

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\Phi(\boldsymbol{V},\boldsymbol{U})=\lambda\sum_{i=1}^{C}w_i^{'}\|\boldsymbol{v}_i-\tilde{\boldsymbol{v}}_i\|^2 \quad (5a)$$

where

$$w_i^{'}=\eta w_i+(1-\eta)\tilde{w}_i, w_i=\sum_{j=1}^{N}\mu_{ij}, \tilde{w}_i=\sum_{j=1}^{N}\tilde{\mu}_{ij}. \quad (5b)$$

For easily interpreting the meaning of Definition 2, we specifically illustrate its overall composition in Fig. 2. As explained in Definition 1, $\sum_{i=1}^{C}\|\boldsymbol{v}_i-\tilde{\boldsymbol{v}}_i\|^2$ can be used to measure the approximation degree between the estimated cluster prototypes in the target domain and the given cluster prototypes in the source domain. However, from the perspective of weighted sum, we need to further differentiate the importance of each $\|\boldsymbol{v}_i - \tilde{\boldsymbol{v}}_i\|^2$, i.e., the individual weight w'_i. In our view, it is reasonable that major clusters, i.e., the clusters composed of many individuals, play primary influence to the total measurement. Therefore, the issue is now converted to find the major clusters in the target domain dataset. As we are aware, each column $[\mu_{1j}, \ldots, \mu_{ij}, \ldots, \mu_{Cj}]^{\mathrm{T}}$ in the membership matrix $\boldsymbol{U} \in R^{C\times N}$ in the target domain indicates the potential affiliation of each individual $\boldsymbol{x}_j$ in the target dataset, and the greater the value of μ_{ij}, the more strongly $\boldsymbol{x}_j$ belongs to the estimated cluster i in the target domain. Let us analyse the problem in the other point of view, i.e., each row $[\mu_{i1}, \ldots, \mu_{ij}, \ldots, \mu_{iN}]$ in $\boldsymbol{U}$. It is definite that cluster i necessarily contains plenty of samples if many entries of $[\mu_{i1}, \ldots, \mu_{ij}, \ldots, \mu_{iN}]$ take larger values, and which also equals that $w_i=\sum_{j=1}^{N}\mu_{ij}$ takes great values. Therefore, as indicated in Fig. 2, $w_i=\sum_{j=1}^{N}\mu_{ij}$ can be intuitively used to measure the probability of cluster i that whether it belongs to the major clusters, i.e., whether it consists of many individuals. Moreover, in light of the similarity between the source domain and the target domain in transfer learning, the membership $\tilde{\mu}_{ij}$ of $\boldsymbol{x}_j$ in the target domain to the known cluster prototype $\tilde{\boldsymbol{v}}_i$ in the source domain can be employed as a reference for μ_{ij}, and recursively, $\tilde{w}_i=\sum_{j=1}^{N}\tilde{\mu}_{ij}$ can be adopted as a reference for $w_i=\sum_{j=1}^{N}\mu_{ij}$. Consequently, as shown in Fig. 2, the combination of w_i and $\tilde{w}_i$ with a trade-off factor $\eta \in [0, 1]$ is devised in (5). Here, the value of η controls the reference degree with respect to $\tilde{\mu}_{ij}$, i.e., $\eta \to 1$ implies that the contribution of μ_{ij} dominates the optimization formulation $\Theta(\boldsymbol{V}, \boldsymbol{U})$, whereas $\eta \to 0$ indicates the membership $\tilde{\mu}_{ij}$ is significantly emphasized. By the way, in light of $\sum_{i=1}^{C}\mu_{ij}=1, \sum_{i=1}^{C}w_i=\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}=\sum_{j=1}^{N}\sum_{i=1}^{C}\mu_{ij}=N$, and $\sum_{i=1}^{C}\tilde{w}_i=N$, thus the sum of all the weights in $\Theta(\boldsymbol{V}, \boldsymbol{U})$, i.e., $\sum_{i=1}^{C}w_i^{'}=\eta\sum_{i=1}^{C}w_i+(1-\eta)\sum_{i=1}^{C}\tilde{w}_i$, equals N rather than 1 in our case.

Both the known cluster prototypes and their associated fuzzy memberships in the source domain are simultaneously utilized in (5), therefore the designed formulation is named as CPM-JL transfer optimization.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

B. Novel CDMEC Framework and Matching Algorithm

In the classical MEC algorithm [see (1)], the term $\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2$ is utilized to evaluate the total deviation of all patterns to all cluster prototypes. Obviously, it is also in the form of weighted sum and the membership μ_{ij} is employed as the weight factor. Inspired by (5), we introduce the transfer trick into this term, and it can be represented as

$$\eta\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2+(1-\eta)\sum_{i=1}^{C}\sum_{j=1}^{N}\tilde{\mu}_{ij}\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2 \quad (6)$$

where parameter η and membership $\tilde{\mu_{ij}}$ are the same as those in (5).

1) Specific CPM-JL Framework for CDMEC—Based on (5) and (6), we herein present a novel, integrated framework for solving the general CDMEC problem.

Definition 3: Let $X_T=\boldsymbol{x}_i|\boldsymbol{x}_i\in R^d, i=1,2,\ldots,N\}$ denote the target domain where d is the data dimensionality and N is the data size. Assume there are C clusters in both the source domain and the target domain. The notations of $\boldsymbol{v}_i$, $\tilde{\boldsymbol{v}}_i$, μ_{ij}, $\tilde{\mu_{ij}}$, η, and λ are the same as those in (5) or (6). Make $\gamma>0$ be the regularization parameter of the Shannon's entropy. In addition, have $\eta\in[0,1]$ and $\lambda\geq 0$, $\lambda\neq=1$, then the CPM-JL framework for CDMEC can be defined as

$$\min_{\boldsymbol{V},\boldsymbol{U}}\left(\begin{array}{c} J_{\mathrm{CPM-JL-CDMEC}}(\boldsymbol{V},\boldsymbol{U})=\eta\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2 \\ +\gamma\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}\ln\mu_{ij} \\ +(1-\eta)\sum_{i=1}^{C}\sum_{j=1}^{N}\tilde{\mu}_{ij}\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2 \\ +\lambda\sum_{i=1}^{C}\sum_{j=1}^{N}\left(\eta\mu_{ij}+(1-\eta)\tilde{\mu}_{ij}\right)\|\boldsymbol{v}_i-\tilde{\boldsymbol{v}}_i\|^2 \end{array}\right) \quad (7)$$
$$\text{s. t.}\quad \mu_{ij}\in[0,1],\ 1\leq i\leq C,\ 1\leq j\leq N,\ \sum_{i=1}^{C}\mu_{ij}=1.$$

There are four items in (7). Both the first and the second terms are inherited from the classical MEC algorithm with only introducing the transfer trade-off coefficient η into the first one. The last two terms, as presented in (5) or (6), are introduced to incorporate additional knowledge from transfer learning.

Theorem 1: The necessary conditions for minimizing the objective function $J_{\mathrm{CPM-JL-CDMEC}}$ in (7) yields the following cluster prototype and membership update equations:

$$\boldsymbol{v}_i=\frac{\sum_{j=1}^{N}\left(\eta\mu_{ij}+(1-\eta)\tilde{\mu}_{ij}\right)\boldsymbol{x}_j-\lambda\sum_{j=1}^{N}\left(\eta\mu_{ij}+(1-\eta)\tilde{\mu}_{ij}\right)\tilde{\boldsymbol{v}}_i}{\sum_{j=1}^{N}\left(\eta\mu_{ij}+(1-\eta)\tilde{\mu}_{ij}\right)-\lambda\sum_{j=1}^{N}\left(\eta\mu_{ij}+(1-\eta)\tilde{\mu}_{ij}\right)} \quad (8)$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\mu_{ij}=\frac{\exp\left(-\frac{\eta\left(\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2+\lambda\|\boldsymbol{v}_i-\tilde{\boldsymbol{v}}_i\|^2\right)}{\gamma}\right)}{\sum_{k=1}^{C}\exp\left(-\frac{\eta\left(\|\boldsymbol{x}_j-\boldsymbol{v}_k\|^2+\lambda\|\boldsymbol{v}_k-\tilde{\boldsymbol{v}}_k\|^2\right)}{\gamma}\right)}. \quad (9)$$

Proof: Using the Lagrange optimization, the minimization of $J_{\text{CPM-JL-CDMEC}}$ can be transformed into the following unconstrained minimization problem:

$$\begin{aligned}L(\mu_{ij},\boldsymbol{v}_i,\boldsymbol{\alpha}_j)=&\eta\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2+\gamma\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}\ln\mu_{ij}\\&+(1-\eta)\sum_{i=1}^{C}\sum_{j=1}^{N}\tilde{\mu}_{ij}\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2\\&+\lambda\sum_{i=1}^{C}\sum_{j=1}^{N}\left(\eta\mu_{ij}+(1-\eta)\tilde{\mu}_{ij}\right)\|\boldsymbol{v}_i-\tilde{\boldsymbol{v}}_i\|^2\\&+\sum_{j=1}^{N}\alpha_j\left(1-\sum_{i=1}^{C}\mu_{ij}\right)\end{aligned} \quad (10)$$

where $\alpha_j (j = 1, \ldots, C)$ are the Lagrange multipliers.

By separately setting the derivatives to zero with respect to μ_{ij} and $\boldsymbol{v}_i$, we arrive at

$$\begin{aligned}&\frac{\partial L}{\partial\mu_{ij}}=\eta\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2+\gamma(1+\ln\mu_{ij})+\lambda\eta\|\boldsymbol{v}_i-\tilde{\boldsymbol{v}}_i\|^2-\alpha_j=0\\&\iff\gamma\ln\mu_{ij}=\alpha_j-\eta\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2-\gamma-\lambda\eta\|\boldsymbol{v}_i-\tilde{\boldsymbol{v}}_i\|^2\\&\iff\mu_{ij}=\exp((\alpha_j-\gamma)/\gamma)\exp\left(\left(-\eta\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2-\lambda\eta\|\boldsymbol{v}_i-\tilde{\boldsymbol{v}}_i\|^2\right)/\gamma\right).\end{aligned} \quad (11)$$

As $\sum_{k=1}^{C}\mu_{kj}=1$, based on (11), we get exp $((\alpha_j-\gamma)/\gamma)$

$$\exp\left((\alpha_j-\gamma)/\gamma\right)=\left(1/\sum_{k=1}^{C}\exp\left(\left(-\eta\|\boldsymbol{x}_j-\boldsymbol{v}_k\|^2-\lambda\eta\|\boldsymbol{v}_k-\tilde{\boldsymbol{v}}_k\|^2\right)/\gamma\right)\right) \quad (12)$$

and then by substituting (12) into (11), we can attain (9).

Algorithm 1

CPM-JL-CDMEC

Inputs:	The target domain dataset X_T, the cluster number C, the maximum iteration number *maxiter*, the iteration termination threshold ε, the known cluster prototypes $\tilde{\boldsymbol{V}}$ or the source domain dataset X_S, the parameter values in Eq. (7), such as η, λ, and γ.
Outputs:	The memberships $\boldsymbol{U}$, the cluster prototypes $\boldsymbol{V}$.
Extracting knowledge in the source domain:	
Setp1:	Obtain the cluster prototypes in $\tilde{\boldsymbol{V}}$ X_S by the classical MEC algorithm; (Skip this step if the cluster prototypes $\tilde{\boldsymbol{V}}$ in X_S are known beforehand)
Step2:	Calculate the reference memberships $\tilde{\boldsymbol{U}}$ of all patterns in X_T to the cluster prototypes $\tilde{\boldsymbol{V}}$ in X_S via Eq. (3);
Cross-domain MEC in the target domain:	

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Step 1:	Set the iteration counter $t = 1$ and initialize the memberships $\boldsymbol{U}(t)$;
Step 2:	Calculate the cluster prototypes $\boldsymbol{V}(t)$ via Eq. (8), $\boldsymbol{U}(t)$, $\tilde{\boldsymbol{U}}$, and $\tilde{\boldsymbol{V}}$;
Step 3:	Calculate the memberships $\boldsymbol{U}(t+1)$ via Eq. (9), $\boldsymbol{V}(t)$ and $\tilde{\boldsymbol{V}}$
Step 4:	If $\|\boldsymbol{U}(t+1) - \boldsymbol{U}(t)\| < \varepsilon$ or $t = maxiter$ go to Step 5, otherwise, $t = t + 1$ and go to Step 2;
Step 5:	Output the final cluster prototypes $\boldsymbol{V}$ and memberships $\boldsymbol{U}$ in the target domain.

$$\frac{\partial L}{\partial \boldsymbol{v}_i} = \sum_{j=1}^{N} (\eta \mu_{ij} + (1-\eta)\,\tilde{\mu}_{ij})\,(\boldsymbol{x}_j - \boldsymbol{v}_i) + \lambda \sum_{j=1}^{N} (\eta \mu_{ij} + (1-\eta)\,\tilde{\mu}_{ij})\,(\boldsymbol{v}_i - \tilde{\boldsymbol{v}}_i) = 0. \quad (13)$$

We can subsequently achieve (8) by rearranging (13).

2) CPM-JL-CDMEC Algorithm—Based on (7)–(9), we now present the core algorithm for CDMEC problems, i.e., CPM-JL-CDMEC, as specifically described in Algorithm 1.

CPM-JL-CDMEC relies essentially on the knowledge coming from the source domain, therefore, the overall workflow of CPM-JL-CDMEC can be divided into two phases: extracting knowledge from the source domain and CDMEC in the target domain.

C. Convergence of CPM-JL-CDMEC

It is well known that the Zangwill's convergence theorem [9], [64], [65] can be adopted for the convergence proof for almost all iterative optimization algorithms. We aim to prove the convergence of CPM-JL-CDMEC by demonstrating it is a special case of Zangwill's theorem that can be summarized as follows.

Lemma 1 (Zangwill's Convergence Theorem)—Let V be a domain of a certain continuous function g, and $S \subset \boldsymbol{V}$ be the solution set of g. Define $A : V \to P(V)$ is a point to set mapping, which creates an iterative sequence $\{\boldsymbol{z}^{(t)} = A^{(t)}(\boldsymbol{z}^{(0)})\}$, where $\boldsymbol{z}^{(0)} \in V$. If the following conditions hold:

1. $\{\boldsymbol{z}^{(t)}\} \subset \Gamma \subseteq V$, Γ is a compact set;
2. the function $g : V \to R$, satisfying:
   a. if $\boldsymbol{z} \notin S$, then for any $\boldsymbol{y} \in A(\boldsymbol{z})$, $g(\boldsymbol{y}) < g(\boldsymbol{z})$;
   b. if $\boldsymbol{z} \in S$, then either the algorithm terminates or for any $\boldsymbol{y} \in A(\boldsymbol{z})$, $g(\boldsymbol{y}) = g(\boldsymbol{z})$.
3. The map A is closed at $\boldsymbol{z}$ if $\boldsymbol{z} \notin S$.

Then, either the algorithm stops at a solution or the limit of any converged subsequence is a solution.

Now, we need to prove that our CPM-JL-CDMEC satisfies the three conditions in Lemma 1. Using the similar strategy proposed in [65], we can prove the corresponding Theorems 2a–2c with the auxiliary Definitions 4–7. In light of limited space, we omit the proof details of Theorems 2a–2c, please refer to [65] for detailed derivations.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Definition 4: Let $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ be a finite dataset in the Euclidean space R^d, then the set of all fuzzy C-partition of X is defined as

$$M_C = \{\boldsymbol{U} \in R^{C\times N} | 0 \le u_{ij} \le 1,\ 1 \le i \le C,\ 1 \le j \le N, \sum_{i=1}^{C} u_{ij} = 1\}.$$

Definition 5: Define $G_1 : M_C \to R^{d\times C}$ to be a function: $G_1(\boldsymbol{U}) = \boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_C)$, where $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{id})^{\mathrm{T}} \in R^d$, $1 \le i \le C$, are calculated according to (8) and $\boldsymbol{U} \in Mc$.

Definition 6: Define $G_2 : R^{d\times C} \to P(M_C)$ to be a point-to-set map: $G_2(\boldsymbol{V}) = \{\boldsymbol{U} | \boldsymbol{U} \in Mc$ and it is computed by (9)$\}$.

Definition 7: Define a point-to-set map $T_{\mathrm{CPM-JL-CDMEC}} : R^{d\times C} \times M_C \to P(R^{d\times C} \times M_C)$ for the iteration in CPM-JL-CDMEC: $T_{\mathrm{CPM-JL-CDMEC}}(\boldsymbol{V},\boldsymbol{U}) = \{(\boldsymbol{V}, \hat{\boldsymbol{U}}) | \boldsymbol{V} = G_1(\boldsymbol{U}), \hat{\boldsymbol{U}} \in G_2(\boldsymbol{V})\}$. Specifically, such map can be rewritten as the following composition $T_{\mathrm{CPM-JL-CDMEC}} = A_2 \circ A_1$, where $A_1(\boldsymbol{V},\boldsymbol{U}) = (G_1(\boldsymbol{U}),\boldsymbol{U})$ and $A_2(\boldsymbol{V}, \boldsymbol{U}) = \{(\boldsymbol{V}, \boldsymbol{U}) | \boldsymbol{U} \in G_2(\boldsymbol{V})\}$. Thus, $(\boldsymbol{V}, \hat{\boldsymbol{U}}) \in T_{\mathrm{CPM-JL-CDMEC}}(\boldsymbol{V}, \boldsymbol{U}) = A_2 \circ A_1(\boldsymbol{V}, \boldsymbol{U}) = A_2(G_1(\boldsymbol{U}),\boldsymbol{U}) = A_2(\boldsymbol{V}, \hat{\boldsymbol{U}}) = \{(\boldsymbol{V}, \hat{\boldsymbol{U}}) | \hat{\boldsymbol{U}} \in G_2(\boldsymbol{V})\}$.

Theorem 2a: Let $\eta \in [0, 1]$, $\lambda \ge 0$ and $\lambda \ne 1$, and $\gamma > 0$ take the specific values as well as $\tilde{\boldsymbol{U}}$ and $\tilde{\boldsymbol{V}}$ be fixed, suppose $X = \{x_1, \ldots, x_N\}$ contains at least C ($C < N$) distinct points, and make the solution set S of the optimization problem $\min_{(\boldsymbol{V},\boldsymbol{U}) \in R^{d\times C}\times M_C} J_{\mathrm{CPM-JL-CDMEC}}(\boldsymbol{V},\boldsymbol{U})$ be defined as

$$S = \left\{ (\overline{\boldsymbol{V}}, \overline{\boldsymbol{U}}) \in R^{d\times C} \times M_C \middle| \begin{array}{c} J_{\mathrm{CPM-JL-CDMEC}}\left(\overline{\boldsymbol{V}}, \overline{\boldsymbol{U}}\right) \\ \le J_{\mathrm{CPM-JL-CDMEC}}\left(\overline{\boldsymbol{V}}, \boldsymbol{U}\right), \\ \forall \boldsymbol{U} \in M_C \text{ and} \\ J_{\mathrm{CPM-JL-CDMEC}}\left(\overline{\boldsymbol{V}}, \overline{\boldsymbol{U}}\right) \\ < J_{\mathrm{CPM-JL-CDMEC}}\left(\boldsymbol{V}, \overline{\boldsymbol{U}}\right), \\ \forall \boldsymbol{V} \ne \overline{\boldsymbol{V}}. \end{array} \right\} \quad (14)$$

Then, it holds that $J_{\mathrm{CPM-JL-CDMEC}}(\boldsymbol{V}, \hat{\boldsymbol{U}}) \le J_{\mathrm{CPM-JL-CDMEC}}(\boldsymbol{V},\tilde{\boldsymbol{U}})$ for each $(\boldsymbol{V}, \hat{\boldsymbol{U}}) \in T_{\mathrm{CPM-JL-CDMEC}}(\boldsymbol{V},\tilde{\boldsymbol{U}})$, and the inequality is strict if $(\boldsymbol{V},\tilde{\boldsymbol{U}}) \notin S$.

Theorem 2b: Suppose $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ contains at least C($C < N$) distinct points and $(\boldsymbol{V}^{(0)},\boldsymbol{U}^{(0)})$ is the starting point of iteration of $T_{\mathrm{CPM-JL-CDMEC}}$ with $\boldsymbol{U}^{(0)} \in M_C$ and $\boldsymbol{V}^{(0)} = G_1(\boldsymbol{U}^{(0)})$, then the iteration sequence $\{\boldsymbol{V}^{(t)}, (\boldsymbol{U}^{(t)})\}$, $t = 1, 2, \ldots$, is contained in a compact subset of $R^{d\times C} \times M_C$.

Theorem 2c: Let $\eta \in [0, 1]$, $\lambda \ge 0$ and $\lambda \ne 1$, and $\gamma > 0$ take the specific values as well as $\tilde{\boldsymbol{U}}$ and $\tilde{\boldsymbol{V}}$ be fixed, suppose $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ contains at least C($C < N$) distinct points, then the point-to-set map $T_{\mathrm{CPM-JL-CDMEC}} : R^{d\times C} \times M_C \to P(R^{d\times C} \times M_C)$ is closed at each point in $R_{d\times C} \times M_C$.

As the objective function $J_{\mathrm{CPM-JL-CDMEC}}$ defined in (7) is continuous, the convergence of the CPM-JL-CDMEC algorithm, as summarized in following Theorem 3, is immediately guaranteed from Theorems 2a–2c and Lemma 1.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Theorem 3: (Convergence of CPM-JL-CDMEC). Let $\eta \in [0, 1]$, $\lambda \geq 0$ and $\lambda \neq 1$, and $\gamma > 0$ take the specific values as well as $\tilde{\boldsymbol{U}}$ and $\tilde{\boldsymbol{V}}$ be fixed, have $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ contain at least $C (C < N)$ distinct points and $J_{\text{CPM–JL–CDMEC}}(\boldsymbol{V}, \boldsymbol{U})$ be defined in (7), make $(\boldsymbol{V}^{(0)}, \boldsymbol{U}^{(0)})$ be the starting point of the iterations via $T_{\text{CPM–JL–CDMEC}}$ with $\boldsymbol{U}^{(0)} \in M_C$ and $\boldsymbol{V}^{(0)} = G_1(\boldsymbol{U}^{(0)})$, then the iteration sequence $\{\boldsymbol{V}^{(t)}, (\boldsymbol{U}^{(t)})\}$, $t = 1, 2, \ldots$, either terminates at a point $(\boldsymbol{V}^*, \boldsymbol{U}^*)$ in the solution set S or there is a subsequence converging to a point in S.

D. Parameter Setting

There are three important parameters in CPM-JL-CDMEC, i.e., the Shannon's entropy parameter γ, the transfer regularization coefficient λ, and the trade-off factor η. As usual, the well-known grid search strategy is used for parameter setting in this paper. As we are aware, the grid search strategy depends on some validity indices and current validity indices can be divided roughly into two categories: external criteria (label-based) and internal criteria (label-free) [66]. Those existing external indices, such as normalized mutual information (NMI) [23], [42], [67] and rand index (RI) [67], or internal ones, e.g., Davies–Bouldin index (DBI) [68] and Dunn index [68], can be certainly deployed for the grid search process in this paper. However, for our specific CPM-JL-CDMEC algorithm, especially when it runs on real-life datasets, it is necessary that a dedicated validity index is developed for further enhancing its practicability. For this purpose, we establish the following new validity index.

Definition 8—Let $\tilde{Q} = \{\tilde{\boldsymbol{x}}_l\} (l = 1, \ldots, m)$ denote the subset constructed by extracting the $p_i (i = 1, \ldots, C)$ nearest samples to each cluster prototype $\tilde{\boldsymbol{v}}_i (i = 1, \ldots, C)$ in the source domain, where $m = \sum_{i=1}^{C} p_i$ denotes the total data size in $\tilde{Q}$. Let $Q = \{\boldsymbol{x}_j\} (j = 1, \ldots, m)$ signify the subset constructed by extracting one of the nearest neighbors in the target domain to each sample in $\tilde{Q}$. Furthermore, have $\boldsymbol{U}_{C\times m}^{\text{SS}} = [\boldsymbol{\mu}_1^{\text{SS}}, \ldots, \boldsymbol{\mu}_m^{\text{SS}}]$ denote the membership matrix where $\boldsymbol{\mu}_l^{\text{SS}} = [\mu_{1l}^{\text{SS}}, \ldots, \mu_{Cl}^{\text{SS}}]^{\text{T}}$, $l = 1, \ldots, m$, represents the fuzzy membership vector of each sample $\tilde{\boldsymbol{x}}_l$ in $\tilde{Q}$ to the known cluster prototypes $\tilde{\boldsymbol{V}}$ in the source domain; $\boldsymbol{U}_{C\times m}^{\text{ST}} = [\boldsymbol{\mu}_1^{\text{ST}}, \ldots, \boldsymbol{\mu}_m^{\text{ST}}]$ denote the membership matrix where $\boldsymbol{\mu}_j^{\text{ST}} = [[\mu_{1j}^{\text{ST}}, \ldots, \mu_{Cj}^{\text{ST}}]^{\text{T}}$, $j = 1, \ldots, m$, represents the fuzzy membership vector of each sample $\boldsymbol{x}_j$ in Q to the known cluster prototypes $\tilde{\boldsymbol{V}}$ in the source domain; and $\boldsymbol{U}_{C\times m}^{\text{TT}} = [\boldsymbol{\mu}_1^{\text{TT}}, \ldots, \boldsymbol{\mu}_{C_j}^{\text{TT}}]$ denote the membership matrix where $\boldsymbol{\mu}_j^{\text{TT}} = [\mu_{1j}^{\text{TT}}, \ldots, \mu_{C_j}^{\text{TT}}]^{\text{T}}$, $j = 1, \ldots, m$, signifies the fuzzy membership vector of each sample $\boldsymbol{x}_j$ in Q to the estimated cluster prototypes $\boldsymbol{V}$ in the target domain, respectively. Thus, the new validity index, named FM-CDDM, for our proposed CPM-JL-CDMEC can be defined as

$$\text{FM- CDDM} = \left| \frac{\sum_{k=1}^{m} \|\boldsymbol{\mu}_k^{\text{SS}} - \boldsymbol{\mu}_k^{\text{ST}}\|}{m} - \frac{\sum_{k=1}^{m} \|\boldsymbol{\mu}_k^{\text{TT}} - \boldsymbol{\mu}_k^{\text{ST}}\|}{m} \right|. \quad (15)$$

FM-CDDM, in the form of (15), is designed according to this idea: to measure the differences between the source domain and the target domain, the average distances of each pair of membership vectors in "$\boldsymbol{U}_{C\times m}^{\text{SS}}$ versus $\boldsymbol{U}_{C\times m}^{\text{ST}}$" and "$\boldsymbol{U}_{C\times m}^{\text{TT}}$ versus $\boldsymbol{U}_{C\times m}^{\text{ST}}$" are

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

separately employed and calculated. Actually, the individuals in subset $\tilde{Q}$ are the representatives of each cluster in the source domain, and those in subset Q in the target domain can be regarded as the most similar variations of the corresponding samples in $\tilde{Q}$. Hence, we aim to utilize these two subsets to estimate the differences between the source domain and the target domain from two perspectives. One is that, we fix the cluster prototypes and vary the data samples, i.e., we first separately compute the fuzzy memberships $\boldsymbol{U}_{C\times m}^{\mathrm{SS}}$ and $\boldsymbol{U}_{C\times m}^{\mathrm{ST}}$ of $\tilde{Q}$ and Q to the same known cluster prototypes $\tilde{\boldsymbol{V}}$ in the source domain, and then figure out the average distance via the first term in (15) as one estimated difference. The other is converse, i.e., we adopt the same data samples but different cluster prototypes. Specifically, we firstly compute the fuzzy memberships $\boldsymbol{U}_{C\times m}^{\mathrm{ST}}$ and $\boldsymbol{U}_{C\times m}^{\mathrm{TT}}$ of Q to the known cluster prototypes $\tilde{\boldsymbol{V}}$ in the source domain and the estimated cluster prototypes $\boldsymbol{V}$ in the target domain, respectively, and then calculate the average distance via the second term in (15) as the other estimated difference.

As a result, the best settings of all the parameters η, λ, and γ involved in (15), in principle, can be simultaneously determined via the grid search strategy when FM-CDDM achieves the smallest value. It should be pointed out that, however, the sample size m of $\tilde{Q}$ and Q influences the effectiveness of FM-CDDM to a certain extent. According to our extensive empirical studies, it is an acceptable setting with $p_i=\min\{120,|\mathrm{Cluster}_i^T|\times 0.5\}$ and $m=\sum_{i=1}^{C}p_i$, where C is the cluster number and $|\mathrm{Cluster}_i^T|$ denotes the data capacity of the ith cluster in the target domain.

IV. Experimental Results

A. Setup

In this section, we focus on evaluating the performance of the developed CPM-JL-CDMEC algorithm and the FM-CDDM index. Besides CPM-JL-CDMEC and MEC, we employ four other algorithms for comparison, i.e., learning shared subspace for multitask clustering (LSSMTC) [48], combining K-means (CombKM) [48], self-taught clustering (STC) [61], and transfer spectral clustering (TSC) [45]. These algorithms are good representatives of the state-of-the-art algorithms related to our studies. MEC and CPM-JL-CDMEC belong to soft partition clustering; CombKM and LSSMTC belong to hard partition clustering; LSSMTC, TSC, and CombKM belong to multitask clustering; STC, TSC, and CPM-JL-CDMEC belong to transfer clustering. In addition, STC and TSC also belong essentially to co-clustering.

Our experiments were implemented on both synthetic and real-life datasets. To verify the clustering performance of these involved algorithms, three popular validity indices were adopted in this paper, i.e., NMI, RI, and DBI. Among them, NMI and RI belong to external criteria, whereas DBI is an internal criterion. In addition, for CPM-JL-CDMEC, the dedicated FM-CDDM index, proposed by ourselves in (15), was also calculated. Next, we briefly review the particular definitions of NMI, RI, and DBI as follows.

1) NMI—

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\mathrm{NMI}=\frac{\sum_{i=1}^{k}\sum_{j=1}^{c}N_{i,j}\log\left(\frac{N\cdot N_{i,j}}{N_i\cdot N_j}\right)}{\sqrt{\left(\sum_{i=1}^{k}N_i\log\frac{N_i}{N}\right)\left(\sum_{j=1}^{c}N_j\log\frac{N_j}{N}\right)}} \quad (16)$$

where $N_{i,j}$ denotes the number of agreements between cluster i and class j, N_i is the number of data points in cluster i, N_j is the number of data points in class j, and N is the size of the whole dataset.

2) RI—

$$\mathrm{RI}=\frac{f_{00}+f_{11}}{N(N-1)/2} \quad (17)$$

where f_{00} denotes the number of any two sample points belonging to two different clusters, f_{11} denotes the number of any two sample points belonging to the same cluster, and N is the total number of sample points.

3) DBI—

$$\mathrm{DBI}=\frac{1}{C}\sum_{k=1}^{C}\max_{k'\neq k}\frac{\delta_k+\delta_{k'}}{\Delta_{kk'}} \quad (18a)$$

where

$$\delta_k=\frac{1}{n_k}\sum_{\boldsymbol{x}_j^k\in C_k}\|\boldsymbol{x}_j^k-\boldsymbol{v}_k\|,\ \Delta_{kk'}=\|\boldsymbol{v}_k-\boldsymbol{v}_{k'}\| \quad (18b)$$

and C denotes the cluster number in the dataset, $\boldsymbol{x}_j^k$ denotes the data point belonging to cluster C_k, n_k, and $\boldsymbol{v}_k$ denote the data size and the centroid of cluster C_k separately.

Both NMI and RI take values within the interval from 0 to 1. Larger values of NMI and RI indicate better clustering performance. In contrast, smaller values of DBI are preferred, and which mean that the levels of both intercluster separation and intracluster compactness are concurrently high. Nevertheless, similar to other internal criteria, DBI has the potential drawback that good values do not necessarily imply better information retrievals. As for FM-CDDM, as previously analysed, smaller values of FM-CDDM are also preferred.

The grid search strategy was adopted in our experiments for parameter optimization. The values or the trial intervals of the primary parameters in each algorithm are listed in Table I.

The experimental results are reported in terms of the means and the standard deviations of the employed validity indices, and which were calculated after 20 repeated runs of each algorithm on each dataset. All experiments were carried out on a personal computer with Intel Core i3-3240 3.4 GHz CPU and 4 GB RAM, Microsoft Windows 7, and MATLAB 2010a.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

B. On Synthetic Datasets

In order to simulate the data scenes for transfer clustering, we generated four synthetic datasets: $\tilde{X}$, X_1, X_2, and X_3. These datasets all contain three base clusters but pose different data distributions, and they were all generated by using the MATLAB built-in function: mvnrnd(). As illustrated in Fig. 3, $\tilde{X}$ simulates the source domain dataset and each cluster consists of 250 samples, so its total capacity is 750.

Let $\boldsymbol{E}_{Ci}$ and $\boldsymbol{\Sigma}_{Ci}$ denote the mean vector and the covariance matrix of the ith cluster in one dataset, respectively. Thus, $\tilde{X}$ was created with $\boldsymbol{E}_{C1}$ = [2 4], $\boldsymbol{E}_{C1}$ = [10 0;0 10], $\boldsymbol{E}_{C2}$ = [9 15], $\boldsymbol{E}_{C2}$ = [25 0;0 7], and $\boldsymbol{E}_{C3}$ = [8 30], $\boldsymbol{E}_{C3}$ = [30 0;0 20]. X_1, X_2, and X_3 are three cases of target domain, as shown in Fig. 4. X_1 simulates the condition where the data in the target domain is insufficient, and which was constructed with $\boldsymbol{E}_{C1}$ = [3 4], $\boldsymbol{E}_{C1}$ = [10 0;0 11], $\boldsymbol{E}_{C2}$ = [10.5 12.5], $\boldsymbol{E}_{C2}$ = [25 0;0 7], $\boldsymbol{E}_{C3}$ = [9 29], $\boldsymbol{E}_{C3}$ = [30 0;0 19.5], and each cluster only consisting of 25 data samples. X_2 and X_3 were generated using the same data distribution as X_1 but each cluster containing 125 samples. Further, two additional clusters were incorporated into X_2 as the interference data with $\boldsymbol{E}_{A1}$ = [7 10], $\boldsymbol{E}_{A1}$ = [4 0;0 4], $\boldsymbol{E}_{A2}$ = [8 20], $\boldsymbol{E}_{A2}$ = [4 0;0 4], and each interference cluster consisting of 35 individuals; and X_3 was added the Gaussian noise with the mean and the deviation being 0 and 2, respectively. In general, the eventual data capacities of $\tilde{X}$, X_1, X_2, and X_3 are separately 750, 75, 445, and 375.

LSSMTC, CombKM, STC, MEC, and CPM-JL-CDMEC were separately carried out on X_1, X_2, and X_3. Among them, except for MEC, the others need to use the source domain $\tilde{X}$ in different manners. Specifically, CPM-JL-CDMEC utilizes the advanced knowledge concluded from $\tilde{X}$, i.e., the known cluster prototypes and the memberships of those patterns in the target domain to these known cluster prototypes in $\tilde{X}$, whereas the others thoroughly use the raw data in $\tilde{X}$. The clustering effectiveness of each algorithm is listed in Table II in terms of the means and the standard deviations of NMI, RI, and DBI, respectively. In addition, in order to validate the reliability of our devised FM-CDDM index, its value was also calculated in CPM-JL-CDMEC with $m=\sum_{i=1}^{C}\min\{120,|\mathrm{Cluster}_i^T|\times 0.5\}$ in (15). Due to the limited space of this paper, we only report the correspondences between FM-CDDM and NMI and RI on X_2 and X_3 in Tables III and IV, respectively.

The TSC algorithm did not run on these synthetic datasets, because it requires that the data dimensionality must be greater than the number of clusters, and which is not met in these synthetic datasets.

Based on the results presented in Tables II–IV, we make some analyses as follows.

1. The data in the target domain X_1 is relatively scarce and clusters 1 and 2 even overlap a bit. In this situation, the clustering outcomes of those approaches working only based on the data itself in the target domain, such as MEC, are prone to inefficient. Furthermore, the data distribution in X_1 is different clearly from that of $\tilde{X}$, so that the clustering performances of LSSMTC, CombKM, and STC are worse than that of CPM-JL-CDMEC. This is because that, even though these competitive clustering algorithms could extract some supporting information from the source

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

domain, the reference value of this information was unreliable in this case. Conversely, CPM-JL-CDMEC utilized the advanced knowledge coming from the source domain as the guidance and such advanced knowledge was more robust than those raw data in this case, therefore it outperformed the others.

2. Although the data capacities in X_2 and X_3 are comparatively sufficient, these two cases were polluted by either the interference data or the noise. In general, one of the merits of transfer learning-based methods is the comparatively high anti-noise capability. Consequently, as we see, CPM-JL-CDMEC and STC performed better than the others on X_2 and X_3.

3. The pursuits of multitask clustering and transfer clustering are different, and which results in the different performance. Multitask clustering focuses on finishing multiple tasks at the same time, and there exist certainly interactivities between these tasks. By comparison, transfer clustering emphasizes the useful information from the source domain which can enhance the learning performance in the target domain. In summary, because of the guidance of transfer information, the performance of transfer clustering approaches, such as CPM-JL-CDMEC and STC, is distinctly better than the others, if we just focus on the clustering results on the target domain datasets.

4. Benefitting from the delicate CPM-JL framework, the clustering effectiveness of CPM-JL-CDMEC is always better than that of MEC. This implies the impact of negative transfer was eliminated in these synthetic scenarios.

5. CPM-JL-CDMEC overcomes the others from the perspective of privacy protection, because CPM-JL-CDMEC merely employs the advanced knowledge (i.e., the known cluster prototypes and their associated fuzzy memberships) in the source domain which cannot be inversely mapped into the original data. Oppositely, the others thoroughly use the raw data in the source domain if they need.

In addition, Tables III and IV indicate that the values of NMI and RI of CPM-JL-CDMEC tend to stay within better ranges when FM-CDDM takes the smallest six values on X_2 and X_3. For example, on X_3, as shown in Table II, the best means of NMI and RI of CPM-JL-CDMEC are 0.7954 and 0.9296, respectively, and in Table IV, the means of NMI and RI are separately around 0.77 and 0.92 when CPM-JL-CDMEC is at the top six. Actually, the correlation between FM-CDDM and NMI/RI exists on all testing datasets, and which indicates that the reliability of FM-CDDM is close to NMI and RI with the distinctive merit that it does not need the sample labels and thus it exhibits stronger practicability.

C. On Real-Life Datasets

In this section, we evaluated the performance of all six algorithms in four real-life transfer scenes, i.e., texture image segmentation, text database, human face recognition, and recognition of real-time human motions (RTHMs). We first introduce the details of these involved real-life datasets, and then present the clustering results of all six algorithms on these datasets.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

1) Constructions of Real Transfer Scenes

a. *Texture image segmentation [datasets: texture data 1 (TD1) and texture data 2 (TD2)]:* For the scenes of texture image segmentation, we chose seven different textures from the *Brodatz texture database*,[1] and based on them, we firstly constructed two original texture images with the same resolution being 100 × 100 = 10 000 pixels, as shown in Fig. 5(a) and (b). To simulate the transfer scenes, we subsequently generated one derivative of each original image by adding noise as shown in Fig. 5(c) and (d) originating separately from Fig. 5(a) and (b). Based on these four texture images, we generated two datasets for transfer clustering: TD1 and TD2. The specific compositions of these two datasets are listed in Table V. Each dataset was obtained by using the Gabor filter method [69] to extract texture features from the corresponding texture images.

b. *Text data clustering [dataset: rec versus talk]:* The New20 text database[2] was used in this paper to construct the transfer scenes for text data clustering. Two categories of text data, i.e., rec and talk, and a few of their sub-categories were employed to deploy the source domain and the target domain. We generated the dataset for our experiment: rec versus talk. The categories and their sub-categories involved in rec versus talk are listed in Table VI. Furthermore, the Bow toolkit [70] was also adopted to reduce the data dimensionality which was originally up to 43 586. The final data used for clustering contains 350 effective features.

c. *Human face recognition [dataset: ORL]:* The widely-used human face repository, i.e., the ORL database of face,[3] was employed for our experiment. We generated the dataset also named ORL in this paper by the following steps. We selected 8 × 10 = 80 facial images from the original ORL database, i.e., eight persons and ten images per person. One facial image of each person is illustrated in Fig. 6. We arbitrarily put eight images per person in the source domain, and the leftover in the target domain. For the purpose of further widening the deviation between the source domain and the target domain as well as enlarging the data capacity in each domain, we also rotated each image anticlockwise with 10° and 20°, respectively. Thus, the eventual image numbers in the source domain and the target domain were separately 192 and 48. In light of the resolution of each image being 92 × 112 = 10304 pixels, we cannot directly use the pixel-gray values of each image as the data features. So we performed the principal component analysis processing on the original gray features, and got the final ORL dataset with data dimensionality being 239.

d. *Recognition of RTHMs [dataset: RTHM]:* The dataset for activities of daily living (ADL) recognition with wrist-worn accelerometer data set in the UCI machine learning repository[4] was adopted in this paper. It was originally composed of numerous three-variate time series which recorded three signal values of the

[1] http://www.ee.oulu.fi/research/imag/texture/image_data/Brodatz32.html
[2] http://www.cs.nyu.edu/~roweis/data.html
[3] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html
[4] http://archive.ics.uci.edu/ml/datasets/

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

sensors worn on 16 volunteers' wrists when they conducted 14 categories of ADL. To simulate the transfer scenario, we first divided the volunteers into two groups according their genders and selected ten categories of activities from the original dataset whose series number are greater than 15, i.e., *climb_stairs*, *comb_hair*, *descend_stairs*, *drink_glass*, *getup_bed*, *liedown_bed*, *pour_water*, *sitdown_chair*, *standup_chair*, and *walk*. In light of the fact that the female's total records are distinctly more than the male's, we used all the female's time series as the source domain and the male's as the target domain. Because the time series dimensions (also, series lengths) of different categories of activities are inconsistent and they vary from hundreds to thousands, the multiscale discrete Haar wavelet decomposition [71] strategy was adopted in this paper for dimensionality reduction. After three to six levels of Haar discrete wavelet transform [71] performed on these raw time series, we truncated the intermediates with the same length being 17 and reshaped them into the forms of vectors, thus, we attained the eventual dataset denoted as RTHM in our experiment with the final data dimension being $17 \times 3 = 51$.

The details of all of the real-life datasets involved in our experiments are shown in Table VII. All these datasets had been normalized before they were adopted in our experiments.

2) Results of the Experiments—Table VIII reports the means and the standard deviations of NMI, RI, and DBI obtained by the six algorithms on all the real-life datasets. These results clearly prove that CPM-JL-CDMEC is generally of the best performance among all of the candidates. It should be mentioned that these real-life datasets are good representatives aside from their different application backgrounds. In details, rec versus talk and ORL both assume high-dimensionality; TD1 and TD2 are of medium-large scales; and RTHM originating from real-time time series belongs to one of hot topics with respect to the applications of intelligent techniques. This implies that, generally speaking, our CPM-JL-CDMEC algorithm can cope effectively with all of these data situations. Moreover, compared with MEC, the average performance improvement of CPM-JL-CDMEC is more than 50% in terms of the NMI index, which further indicates that the CPM-JL framework is an effective way for the CDMEC problem, even if in some complex data scenarios.

Table VIII also reveals two other facts.

1. The LSSMTC and ComKM approaches good at multitask processing are inefficient in high-dimensional data scenarios.
2. The DBI metric, which is calculated based on the internal criterion, cannot always catch the inherent information in the data even if it reaches a good value, such as the value of DBI of STC on ORL. In this regard, our proposed FM-CDDM index demonstrates better versatility as well as practicability. The values of FM-CDDM in CPM-JL-CDMEC on all the real-life datasets were computed during our experiments. Similar to that on the synthetic datasets, the values of NMI and RI in CPM-JL-CDMEC correlate well with FM-CDDM because they really rank near the top when FM-CDDM takes the top six values in all these real-life transfer

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

scenarios. For saving paper length, we only present the cases on three datasets: TD2, ORL, and RTHM, as shown in Tables IX–XI.

The segmentation results of six algorithms on Fig. 5(d) are illustrated in Fig. 7. The pixels with the same labels are displayed in the same colors in each sub-graph corresponding to each algorithm. Intuitively, CPM-JL-CDMEC and TSC achieved the better segmentations.

D. Robustness Analyses

Last but not least, we have evaluated the robustness of our CPM-JL-CDMEC algorithm with respect to its three core parameters, i.e., the transfer trade-off factors η, the transfer regularization parameter λ, and the entropy regularization parameter γ, on both the synthetic and the real-life datasets. On each dataset, we took turns fixing two of the three parameters and gradually changed the third one until CPM-JL-CDMEC achieved the optima by grid search. We calculated the values of the three metrics: NMI, RI, and DBI. Due to the limit of paper length, here, we only report the experimental results on the synthetic dataset X_2, the texture image dataset TD1, and the human motion recognition dataset RTHM.

On X_2, CPM-JL-CDMEC roughly reached the optima with $\lambda = 2$, $\gamma = 100$, and $\eta = 0.8$; on TD1, with $\lambda = 0.06$, $\gamma = 0.01$, and $\eta = 0.9$; and on RTHM, with $\lambda = 0.05$, $\gamma = 0.02$, and $\eta = 0.75$. The performance curves of CPM-JL-CDMEC on these three datasets are illustrated in Fig. 8, where Fig. 8(a)–(c) shows the cases on X_2, Fig. 8(d)–(f) are on TD1, and Fig. 8(g)–(i) are on RTHM.

As revealed in Fig. 8, the clustering effectiveness of CPM-JL-CDMEC is relatively stable when the three major parameters are within proper ranges, which demonstrates that CPM-JL-CDMEC features a good robustness against parameter setting.

V. Conclusion

Inspired by transfer learning, we propose the CPM-JL-CDMEC algorithm as well as the dedicated FM-CDDM index in this paper to deal with cross-domain partition-based clustering issues, especially in the situations where the data could be insufficient or polluted by unknown noise or outliers.

In summary, owing to the delicate, CPM-JL framework and the reliable strategy for avoiding negative transfer issues, CPM-JL-CDMEC proves satisfactory clustering effectiveness and robustness in both the artificial and the real-life transfer scenarios. Besides these, by means of the dedicated FM-CDDM index, another contribution of this paper, the combination of “CPM-JL-CDMEC + FM-CDDM” can cope with most of cross-domain data cases. In addition, the intrinsic mechanism of privacy protection in CPM-JL-CDMEC further strengthens the practicability of our research in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61202311 and Grant 61272210, in part by the Natural Science Foundation of Jiangsu Province under Grant BK201221834, in part by the Research and Development Frontier Grant of Jiangsu Province under Grant BY2013015-02, in part by the Ministry of Education Program for New Century Excellent Talents under Grant NECT-120882, in part by the Fundamental Research Funds for the Central Universities under Grant JUSRP51321B, and in part by the Jiangsu

Province Outstanding Youth Fund under Grant BK20140001. This paper was recommended by Associate Editor A. F. Skarmeta Gomez.

Biographies

Pengjiang Qian (M’12) received the Ph.D. degree from Jiangnan University, Wuxi, China, in 2011.

He is an Associate Professor with the School of Digital Media, Jiangnan University, and Case Western Reserve University, Cleveland, OH, USA, as a Research Scholar in medical image processing. His current research interests include data mining, pattern recognition, bioinformatics and their applications, such as analysis and processing for medical imaging, intelligent traffic dispatching, and advanced business intelligence in logistics. He has authored/co-authored over 30 papers in international/national journals and conferences.

Yizhang Jiang (M’12) is currently pursuing the Ph.D. degree from the School of Digital Media, Jiangnan University, Wuxi, China.

He is currently a Research Assistant with the Department of Computing, Hong Kong Polytechnic University, Hong Kong. His current research interests include pattern recognition, intelligent computation, and their applications. He has published several papers in international journals including the IEEE Transactions on Fuzzy Systems and the IEEE Transactions on Neural Networks and Learning Systems.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

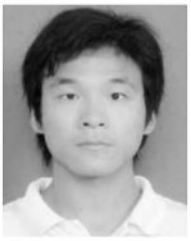
Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Zhaohong Deng (SM'13) received the Ph.D. degree from Jiangnan University, Wuxi, China, in 2008.

He is currently an Associate Professor with the School of Digital Media, Jiangnan University. He visited the University of California, Davis, Davis, CA, USA, and the Hong Kong Polytechnic University, Hong Kong, for over two years. His current research interests include neuro-fuzzy systems, pattern recognition, and their applications. He has authored/co-authored over 40 research papers in international/national journals.

Lingzhi Hu received the Ph.D. degree in medical imaging from Washington University, St. Louis, St. Louis, MO, USA, in 2012.

He is currently a Research Scientist with Philips Electronics North America, Cleveland, OH, USA. His current research interests include system design, image reconstruction, and processing for radiological imaging device. He has authored and co-authored over 30 papers on internationally recognized journals and conferences. He has been invited to review for top journals and conferences in medical imaging for over 20 times.



Shouwei Sun is currently pursuing the M.S. degree from the School of Digital Media, Jiangnan University, Wuxi, China.

His current research interests include pattern recognition, bioinformatics, and their applications. He has authored or co-authored several papers in international/national journals and conferences.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Shitong Wang received the M.S. degree in computer science from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1987.

He visited London University, London, U.K., and Bristol University, Bristol, U.K., the Hiroshima International University, Hiroshima, Japan, Osaka Prefecture University, Osaka, Japan, Hong Kong University of Science and Technology, Hong Kong, and the Hong Kong Polytechnic University, Hong Kong, as a Research Scientist, for over six years. He is currently a Full Professor with the School of Digital Media, Jiangnan University, Wuxi, China. His current research interests include artificial intelligence, neuro-fuzzy systems, pattern recognition, and image processing. He has published over 100 papers in international/national journals and has authored seven books.

Raymond F. Muzic, Jr. (SM’00) received the Ph.D. degree from Case Western Reserve University, Cleveland, OH, USA, in 1991.

He is currently an Associate Professor of Radiology, Biomedical Engineering, and General Medical Sciences—Oncology, Case Western Reserve University. His current research interests include development and application of quantitative methods for medical imaging. He has authored/co-authored approximately 50 peer-reviewed articles. He has led or been a team member on numerous funded research projects. He has also had the pleasure to serve as an advisor for doctoral students.

References

1. MacQueen, JB. Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Stat. Probab; Berkeley, CA, USA. 1967. p. 281-297.
2. Dunn J. A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters. J Cybern. 1974; 3(3):32–57.
3. Höppner, F.; Klawonn, F.; Kruse, R. Fuzzy Cluster Analysis: Methods for Classification, Data Analysis, and Image Recognition. New York, NY, USA: Wiley; 1999.
4. Bezdek J, Hathaway R. Numerical convergence and interpretation of the fuzzy c-shells clustering algorithms. IEEE Trans Neural Netw. Sep; 1992 3(5):787–793. [PubMed: 18276477]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

5. Liu Z, Xu S, Zhang Y, Chen CLP. A multiple-feature and multiple-kernel scene segmentation algorithm for humanoid robot. IEEE Trans Cybern. Nov; 2014 44(11):2232–2240. [PubMed: 25248211]
6. Karayiannis, NB. MECA: Maximum entropy clustering algorithm. Proc. IEEE Int. Conf. Fuzzy Syst; Orlando, FL, USA. 1994. p. 630-635.
7. Li, RP.; Mukaidono, MA. A maximum entropy approach to fuzzy clustering. Proc. IEEE Int. Conf. Fuzzy Syst; Yokohama, Japan. 1995. p. 2227-2232.
8. Bradley PS, Mangasarian QL. K-plane clustering. J Glob Optim. 2000; 16(1):23–32.
9. Wu KL, Yu J, Yang MS. A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality tests. Pattern Recognit Lett. 2005; 26(5):639–652.
10. Bezdek JC. A convergence theorem for the fuzzy ISODATA clustering algorithm. IEEE Trans Pattern Anal Mach Intell. Jan; 1980 PAMI-2(1):1–8. [PubMed: 22499617]
11. de Luis Balaguer MA, Williams CM. Hierarchical modularization of biochemical pathways using fuzzy-c means clustering. IEEE Trans Cybern. Aug; 2013 44(8):1473–1484. [PubMed: 24196983]
12. Jiang Y, et al. Collaborative fuzzy clustering from multiple weighted views. IEEE Trans Cybern. to be published.
13. Heller, KA.; Ghahramani, Z. Bayesian hierarchical clustering. Proc. 22th Int. Conf. Mach. Learn; Bonn, Germany. 2005. p. 297-304.
14. Guha, S.; Rastogi, R.; Shim, K. CURE: An efficient clustering algorithm for large database. Proc. ACM SIGMOD Int. Conf. Manage. Data; Seattle, WA, USA. 1998. p. 73-84.
15. Zhuang X, Huang Y, Palaniappan K, Zhao Y. Gaussian mixture density modeling, decomposition, and applications. IEEE Trans Image Process. Sep; 1996 5(9):1293–1302. [PubMed: 18285218]
16. Figueiredo M, Jain A. Unsupervised learning of finite mixture models. IEEE Trans Pattern Anal Mach Intell. Mar; 2002 24(3):381–396.
17. McLachlan, GJ.; Peel, D. Finite Mixture Models. New York, NY, USA: Wiley; 2000.
18. McLachlan, GJ.; Krishnan, T. The EM Algorithm and Extensions. New York, NY, USA: Wiley; 1997.
19. Wang, W.; Yang, J.; Muntz, R. STING: A statistical information grid approach to spatial data mining. Proc. 23rd Int. Conf. Very Large Data Bases (VLDB); San Francisco, CA, USA. 1997. p. 186-195.
20. Sheikholeslami, G.; Chatterjee, S.; Zhang, A. WaveCluster: A multi-resolution clustering approach for very large spatial databases. Proc. 24th Int. Conf. Very Large Data Bases (VLDB); New York, NY, USA. 1998. p. 428-439.
21. Agrawal, R.; Gehrke, J.; Gunopolos, D. Automatic subspace clustering of high dimensional data for data mining applications. Proc. ACM SIGMOD Int. Conf. Manage. Data; Seattle, WA, USA. 1998. p. 94-105.
22. Shi J, Malik J. Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell. Aug; 2000 22(8):888–905.
23. Sarkar S, Soundararajan P. Supervised learning of large perceptual organization: Graph spectral partitioning and learning automata. IEEE Trans Pattern Anal Mach Intell. May; 2000 22(5):504–525.
24. Qian P, Chung FL, Wang S, Deng Z. Fast graph-based relaxed clustering for large data sets using minimal enclosing ball. IEEE Trans Syst, Man, Cybern B, Cybern. Jun; 2012 42(3):672–687. [PubMed: 22318491]
25. Xu X, Huang Z, Graves D, Pedrycz W. A clustering-based graph Laplacian framework for value function approximation in reinforcement learning. IEEE Trans Cybern. Dec; 2014 44(12):2613–2625. [PubMed: 24802018]
26. Zhang Z, Zheng N, Shi G. Maximum-entropy clustering algorithm and its global convergence analysis. Sci China E, Technol Sci. 2001; 44(1):89–101.
27. Ren S, Wang Y. A proof of the convergence theorem of maximum-entropy clustering algorithm. Sci China F, Inf Sci. 2010; 53(6):1151–1158.
28. Xu, D. PhD dissertation. Dept. Electr. and Comput. Eng., Univ. Florida; Gainesville, FL, USA: 1999. Energy, entropy and information potential for neural computation.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

29. Li RP, Mukaidono M. Gaussian clustering method based on maximum-fuzzy-entropy interpretation. Fuzzy Sets Syst. 1999; 102(2):253–258.
30. Ghorbani M. Maximum entropy-based fuzzy clustering by using L1-norm space. Turk J Math. 2005; 29(1):431–438.
31. Lao, L.; Wu, X.; Cheng, L.; Zhu, X. Maximum weighted entropy clustering algorithm. Proc. IEEE Int. Conf. Netw. Sens. Control; Fort Lauderdale, FL, USA. 2006. p. 1022-1025.
32. Yu J. General C-means clustering model. IEEE Trans Pattern Anal Mach Intell. Aug; 2005 27(8): 1197–1211. [PubMed: 16119260]
33. Yu J, Yang MS. A generalized fuzzy clustering regularization model with optimality tests and model complexity analysis. IEEE Trans Fuzzy Syst. Oct; 2007 15(5):904–915.
34. Wang S, Chung KL, Deng Z, Hu D, Wu H. Robust maximum entropy clustering with its labeling for outliers. Soft Comput. 2006; 10(7):555–563.
35. Zhi X, Fan J, Zhao F. Fuzzy linear discriminant analysis-guided maximum entropy fuzzy clustering algorithm. Pattern Recognit. 2013; 46(6):1604–1615.
36. Karayiannis NB, Zervos N. Entropy-constrained learning vector quantization algorithms and their application in image compression. J Electron Imag. Oct; 2000 9(4):495–508.
37. Li K, Guo Z. Image segmentation with fuzzy clustering based on generalized entropy. J Comput. 2014; 9(7):1678–1683.
38. Li L, Ji H, Gao X. Maximum entropy fuzzy clustering with application to real-time target tracking. Signal Process. 2006; 86(11):3432–3447.
39. Zhu, X.; Ghahramani, Z.; Lafferty, JD. Semi-supervised learning using Gaussian fields and harmonic functions. Proc. 20th Int. Conf. Mach. Learn. (ICML); Washington, DC, USA. 2003. p. 912-919.
40. Zhou, D.; Bousquet, O.; Lal, TN.; Weston, J.; Schölkopf, B. Learning with local and global consistency. Proc. Neural Inf. Process. Syst. (NIPS); Vancouver, BC, Canada. 2004. p. 321-328.
41. Breitenbach, M.; Grudic, GZ. Clustering through ranking on manifolds. Proc. 22nd Int. Conf. Mach. Learn. (ICML); Bonn, Germany. 2005. p. 73-80.
42. Nie F, Xu D, Li X. Initialization independent clustering with actively self-training method. IEEE Trans Syst, Man, Cybern B, Cybern. Feb; 2012 42(1):17–27. [PubMed: 22086542]
43. Dhillon, IS.; Mallela, S.; Modha, DS. Information-theoretic co-clustering. Proc. 9th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min. (KDD); Washington, DC, USA. 2003. p. 89-98.
44. Dhillon, IS. Co-clustering documents and words using bipartite spectral graph partitioning. Proc. 7th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min. (KDD); San Fransisco, CA, USA. 2001. p. 269-274.
45. Jiang, WH.; Chung, FL. Machine Learning and Knowledge Discovery in Databases. Berlin, Germany: Springer; 2012. Transfer spectral clustering; p. 789-803.(LNCS 7524)
46. Caruana R. Multitask learning. Mach Learn. 1997; 28(1):41–75.
47. Ando RK, Zhang T. A framework for learning predictive structures from multiple tasks and unlabeled data. J Mach Learn Res. Dec.2005 6:1817–1853.
48. Gu, Q.; Zhou, J. Learning the shared subspace for multi-task clustering and transductive transfer classification. Proc. 9th Int. Conf. Data Min. (ICDM); Miami, FL, USA. 2009. p. 159-168.
49. Pan J, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. Oct; 2010 22(10): 1345–1359.
50. Tao J, Chung FL, Wang S. On minimum distribution discrepancy support vector machine for domain adaptation. Pattern Recognit. 2012; 45(11):3962–3984.
51. Gao, J.; Fan, W.; Jiang, J.; Han, J. Knowledge transfer via multiple model local structure mapping. Proc. 14th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min. (KDD); Las Vegas, NV, USA. 2008. p. 283-291.
52. Mihalkova, L.; Huynh, T.; Mooney, RJ. Mapping and revising Markov logic networks for transfer learning. Proc. Conf. Artif. Intell. (AAAI); Vancouver, BC, Canada. 2007. p. 608-614.
53. Mihalkova, L.; Mooney, RJ. Transfer learning by mapping with minimal target data. Proc. Conf. Artif. Intell. (AAAI); Chicago, IL, USA. 2008.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

54. Duan L, Tsang IW, Xu D. Domain transfer multiple kernel learning. IEEE Trans Pattern Anal Mach Intell. Mar; 2012 34(3):465–479. [PubMed: 21646679]
55. Yang, P.; Tan, Q.; Ding, Y. Bayesian task-level transfer learning for non-linear regression. Proc. Int. Conf. Comput. Sci. Softw. Eng; Hubei, China. 2008. p. 62-65.
56. Borzemski, L.; Starczewski, G. Application of transfer regression to TCP throughput prediction. Proc. 1st Asian Conf. Intell. Inf. Database Syst; Dong Hoi, Vietnam. 2009. p. 28-33.
57. Mao, W.; Yan, G.; Bai, J.; Li, H. Advances in Neural Networks. Berlin, Germany: Springer; 2010. Regression transfer learning based on principal curve; p. 365-372.(LNCS 6063)
58. Deng Z, Jiang Y, Choi KS, Chung FL, Wang S. Knowledge-leverage-based TSK fuzzy system modeling. IEEE Trans Neural Netw Learn Syst. Aug; 2013 24(8):1200–1212. [PubMed: 24808561]
59. Wang, Z.; Song, YQ.; Zhang, CS. Machine Learning and Knowledge Discovery in Databases. Berlin, Germany: Springer; 2008. Transferred dimensionality reduction; p. 550-565.(LNCS 5212)
60. Pan, SJ.; Kwok, JT.; Yang, Q. Transfer learning via dimensionality reduction. Proc. Conf. Artif. Intell. (AAAI); Chicago, IL, USA. 2008. p. 677-682.
61. Dai, W.; Yang, Q.; Xue, G.; Yu, Y. Self-taught clustering. Proc. 25th Int. Conf. Mach. Learn. (ICML); Helsinki, Finland. 2008. p. 200-207.
62. Gu, Q.; Zhou, J. Transfer heterogeneous unlabeled data for unsupervised clustering. Proc. 21st Int. Conf. Pattern Recognit; Tsukuba, Japan. 2012. p. 1193-1196.
63. Sun, S.; Jiang, Y.; Qian, P. Transfer learning based maximum entropy clustering. Proc. 4th IEEE Int. Conf. Inf. Sci. Technol; Shenzhen, China. 2014. p. 829-832.
64. Zangwill, W. Nonlinear Programming: A Unified Approach. Englewood Cliffs, NJ, USA: Prentice-Hall; 1969.
65. Gan G, Wu J. A convergence theorem for the fuzzy subspace clustering (FSC) algorithm. Pattern Recognit. 2008; 41(6):1939–1947.
66. Liu Y, et al. Understanding and enhancement of internal clustering validation measures. IEEE Trans Cybern. Jun; 2013 43(3):982–993. [PubMed: 23193245]
67. Liu J, et al. Distance-based clustering of CGH data. Bioinformatics. 2006; 22(16):1971–1978. [PubMed: 16705014]
68. Desgraupes, B. Clustering Indices. Univ. Paris Ouest, Lab Modal'X; Nanterre, France: 2013.
69. Kyrki V, Kamarainen JK, Kalviainen H. Simple Gabor feature space for invariant object recognition. Pattern Recognit Lett. 2004; 25(3):311–318.
70. McCallum, AK. Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering [EB/OL]. 1996. [Online]. Available: http://www.cs.cmu.edu/mccallum/bow
71. He X, Shao C, Xiong Y. A new similarity measure based on shape information for invariant with multiple distortions. Neurocomputing. Apr.2014 129:556–569.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

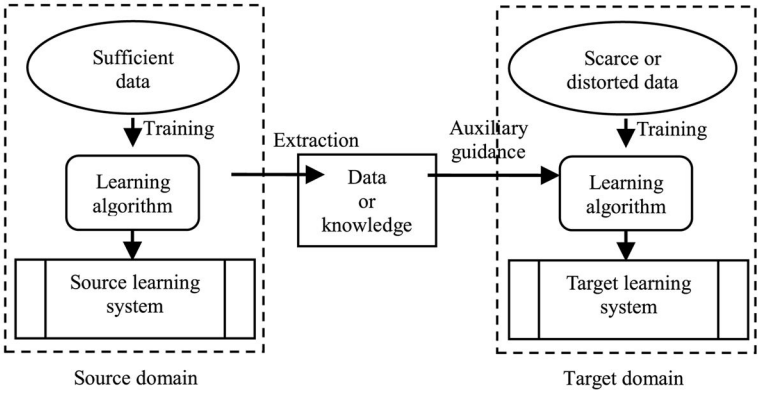


Fig. 1.
Overall framework of transfer learning.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

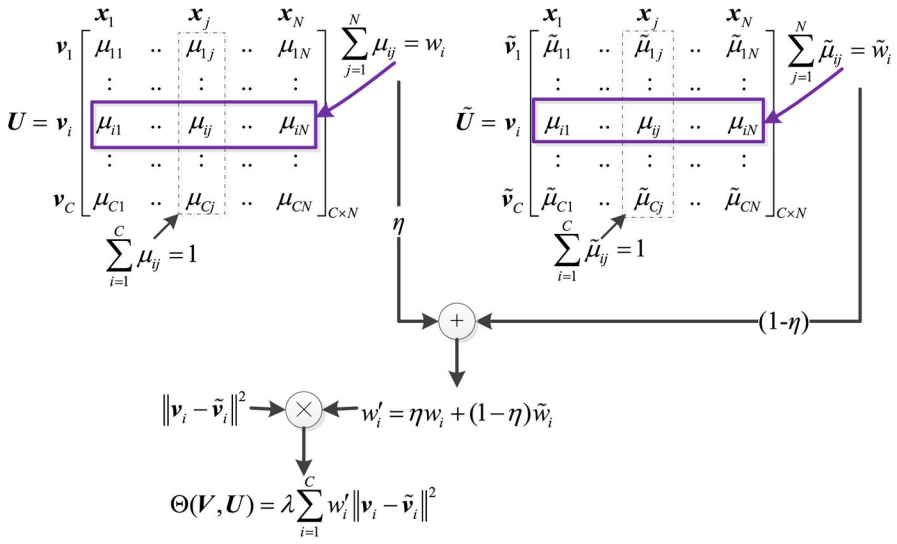


Fig. 2.
Illustration of Definition 2.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

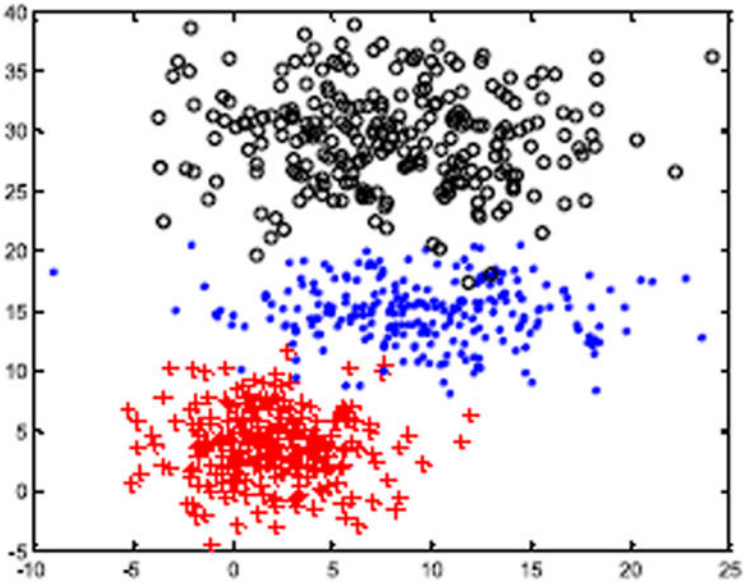


Fig. 3.
Source domain $\tilde{X}$.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

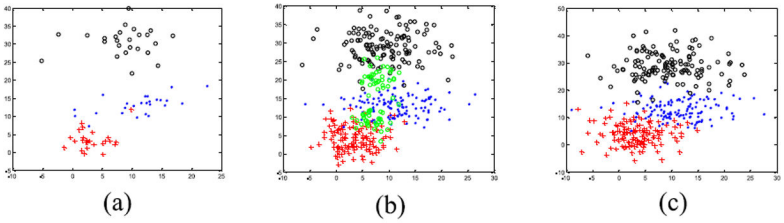


Fig. 4.
Target domains. (a) X_1. (b) X_2. (c) X_3.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

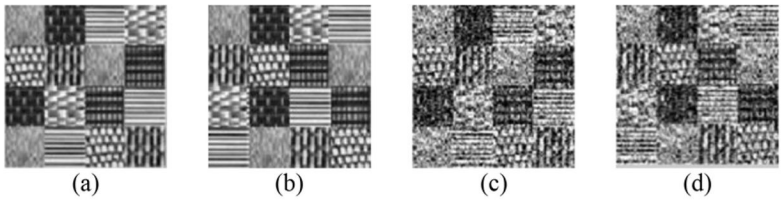


Fig. 5.

Texture images adopted in our experiment. (a) Original texture image 1. (b) Original texture image 2. (c) Derivative image from original image 1. (d) Derivative image from original image 2.

Fig. 6.
Human facial dataset: ORL.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

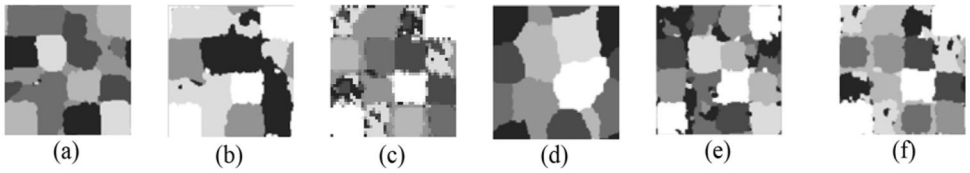


Fig. 7.

Segmentation results of six algorithms on Fig. 5(d). Results of (a) LSSMTC, (b) CombKM, (c) MEC, (d) STC, (e) TSC, and (f) CPM-JL-CDMEC.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

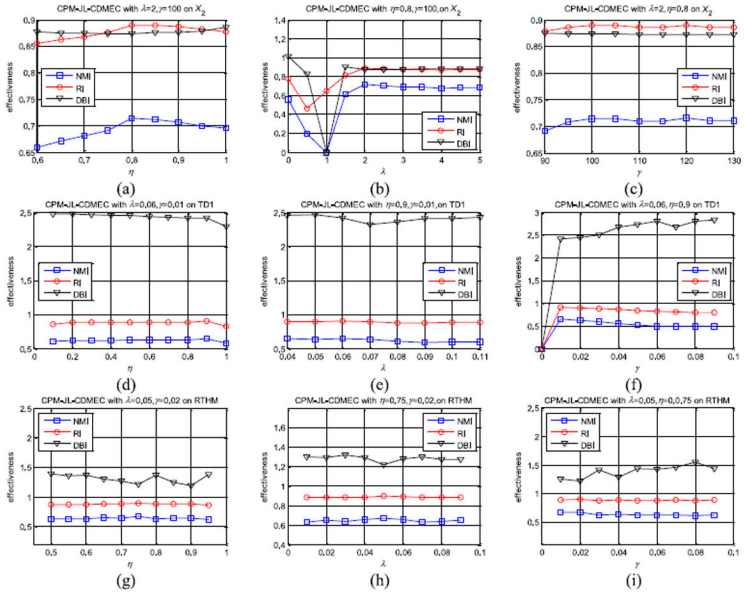


Fig. 8.

Performance curves of CPM-JL-CDMEC with respect to parameters η, λ, and γ on X_2, TD1, and RTHM. (a) On X_2, λ and γ are fixed and η varies. (b) On X_2, η and γ are fixed and λ varies. (c) On X_2, λ and η are fixed and γ varies. (d) On TD1, λ and γ are fixed and η varies. (e) On TD1, η and γ are fixed and λ varies. (f) On TD1, λ and η are fixed and γ varies. (g) On RTHM, λ and γ are fixed and η varies. (h) On RTHM, η and γ are fixed and λ varies. (i) On RTHM, λ and η are fixed and γ varies.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE I

Values or Intervals of Primary Parameters in Six Algorithms

Algorithm	Parameter setting
ComKM	K equals to the number of cluster
LSSMTC	Task number $T = 2$ Regularization parameter $l \in \{2,2^2,2^3,2^4\} \cup \{100:100:1000\}$ Regularization parameter $\lambda \in \{0.25, 0.5, 0.75\}$
STC	Trade-off parameter $\lambda = 1$
TSC	K=27, $\lambda = 3$, and $step = 1$
MEC	Entropy regularization parameter $\gamma \in \{0.01:0.01:0.09\} \cup \{0.1:0.1:1\} \cup \{2:1:10\} \cup \{20:10:100\}$
CPM-JL-CDMEC	Entropy regularization parameter $\gamma \in \{0.01:0.01:0.09\} \cup \{0.1:0.1:1\} \cup \{2:1:10\} \cup \{20:10:100\}$ Transfer regularization parameter $\gamma \in \{0:0.01:0.09\} \cup \{0.1:0.1:0.9\} \cup \{2:1:10\} \cup \{20:10:100\}$ Transfer trade-off factor $\eta \in \{0:0.05:1\}$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE II

Clustering Performance (NMI, RI, and DBI) of Related Algorithms on Synthetic Datasets

Dataset	Metrics	Algorithm				
		LSSMTC	**CombKM**	**MEC**	**STC**	**CPM-JL-CDMEC**
X_1	NMI-mean	0.7626	0.7859	0.7730	0.8552	**0.9144**
	NMI-std	1.17E-16	9.79E-17	1.17E-16	0	0
	RI-mean	0.9023	0.9168	0.9089	0.9443	**0.9654**
	RI-std	1.17E-16	2.34E-16	1.17E-16	0	1.17E-16
	DBI-mean	0.6153	0.6676	0.6229	**0.5928**	0.5929
	DBI-std	0	1.17E-16	1.17E-16	0	0
X_2	NMI-mean	0.6443	0.6960	0.6498	**0.7335**	0.7138
	NMI-std	1.17E-16	0.0036	1.17E-16	0	1.17E-16
	RI-mean	0.8382	0.8738	0.8405	0.8822	**0.8891**
	RI-std	1.17E-16	0.0014	1.17E-16	0	1.17E-16
	DBI-mean	0.8835	0.7931	0.8418	0.8897	**0.6503**
	DBI-std	9.72E-5	0	0.0151	0	0
X_3	NMI-mean	0.7263	0.7163	0.7523	0.7782	**0.7954**
	NMI-std	1.17E-16	1.17E-16	0	0	0
	RI-mean	0.8984	0.8917	0.9111	0.9168	**0.9296**
	RI-std	0	2.34E-16	1.17E-16	0	0
	DBI-mean	0.8820	1.0552	0.8802	0.8806	**0.6605**
	DBI-std	0	0.0119	1.17E-16	0	0

IEEE Trans Cybern. Author manuscript; available in PMC 2017 January 01.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE III

Correspondences Between FM-CDDM and NMI and RI on X_2 (in the Ascending Order of the Means of FM-CDDM)

FM-CDDM-mean	0.03969	0.03985	0.04019	0.04039	0.04046	0.04049
FM-CDDM-std	1.627E-17	1.552E-17	1.151E-17	1.552E-17	9.180E-18	1.097E-17
NMI-mean	0.66410	0.66332	0.66410	0.66846	0.66154	0.66154
NMI-std	0	0	0	0	0	0
RI-mean	0.85444	0.85200	0.85444	0.85942	0.85461	0.85461
RI-std	1.241E-16	0	1.241E-16	1.241E-16	1.241E-16	1.241E-16

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE IV

Correspondences Between FM-CDDM and NMI and RI on X_3 (in the Ascending Order of the Means of FM-CDDM)

FM-CDDM-mean	0.04216	0.04220	0.04223	0.04225	0.04226	0.04227
FM-CDDM-std	0	1.963E-17	1.836E-17	6.939E-18	1.552E-17	4.001E-17
NMI-mean	0.77214	0.77214	0.77214	0.77820	0.77214	0.77820
NMI-std	0	0	0	0	0	0
RI-mean	0.91997	0.91997	0.91997	0.92298	0.91997	0.92298
RI-std	1.241E-16	1.241E-16	1.241E-16	0	1.241E-16	0

IEEE Trans Cybern. Author manuscript; available in PMC 2017 January 01.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE V

Compositions of Texture Datasets Involved in Our Experiment

Dataset	Source domain	Target domain
TD1	Fig. 5(a)	Fig. 5(d)
TD2	Fig. 5(b)	Fig- 5(c)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE VI

Categories and Sub-Categories of New20 Adopted in Our Experiment

Dataset	Source domain	Target domain
rec VS talk	rcc.autos	rec.sport.baseball
	talk.politics.guns	talk.politics.mideast

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE VII

Details of Real-Life Datasets Involved in This Paper

Dataset	Transfer domain	Data size	Dimension	Cluster number
TD1	Source domain	10,000	49	7
	Target domain	10,000	49	
TD2	Source domain	10,000	49	7
	Target domain	10,000	49	
rec VS talk	Source domain	1,500	350	2
	Target domain	500	350	
ORL	Source domain	192	239	8
	Target domain	48	239	
RTHM	Source domain	494	51	10
	Target domain	312	51	

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

IEEE Trans Cybern. Author manuscript; available in PMC 2017 January 01.

TABLE VIII

Clustering Performance (NMI, RI, and DBI) of Different Clustering Algorithms on Real Datasets

Dataset	Metrics	Algorithm					
		LSSMTC	CombKM	MEC	STC	TSC	CPM-JL-CDMEC
TD1	NMI-m	0.5347	0.5153	0.6087	0.4333	0.6198	**0.6489**
	NMI-s	0.0600	0.0828	0.0309	0	0	1.17E-16
	RI-m	0.8010	0.7744	0.8838	0.8304	0.8982	**0.9077**
	RI-s	0.0653	0.1054	0.0147	0	0	2.34E-16
	DBI-m	2.4851	3.0258	2.3648	4.5092	2.8825	**2.3205**
	DBI-s	0.0911	0.0574	0.1139	0	0	0
TD2	NMI-m	0.6004	0.5434	0.5930	0.3421	0.6661	**0.6817**
	NMI-s	0.0838	0.1031	0.0032	0	0	0.0025
	RI-m	0.8375	0.7923	0.8913	0.8042	0.9148	**0.9150**
	RI-s	0.0698	0.0706	0.0031	0	0	0.0015
	DBI-m	2.6604	2.4835	2.7696	4.9633	2.3927	**2.3369**
	DBI-s	0.0859	0.0544	0.1510	0	0	0
rec VS talk	NMI-m	0.0818	0.0572	0.2691	0.1865	0.4224	**0.6470**
	NMI-s	1.46E-17	0.0201	0	0.0055	0	1.17E-16
	RI-m	0.5021	0.5002	0.5960	0.5747	0.7359	**0.8593**
	RI-s	0	0.0004	0	0.0078	0	1.17E-16
	DBI-m	2.2505	2.5320	2.4388	3.9824	1. 7190	**1.4448**
	DBI-s	0	4.68E-16	4.68E-16	0	0	0
ORL	NMI-m	0.3582	0.2124	0.2909	0.3310	0.2950	**0.5218**
	NMI-s	0	0.0954	0	0.0183	0.0054	0
	RI-m	0.7748	0.5870	0.7996	0.8116	0.8124	**0.8667**
	RI-s	0	0.1806	0	0.0034	0.0004	2.34E-16
	DBI-m	5.7024	6.1604	3.2747	**1.5186**	3.2241	3.1124
	DBI-s	0	9.36E-16	0.0934	0.0361	0	0
RTHM	NMI-m	0.4491	0.5120	0.5655	0.5914	0.5986	**0.6737**
	NMI-s	0	0	0.0154	0.0101	0.0040	1.36E-16

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Dataset	Metrics	Algorithm					
		LSSMTC	CombKM	MEC	STC	TSC	CPM-JL-CDMEC
	RI-m	0.7515	0.7745	0.7929	0.8686	0.8862	**0.8985**
	RI-s	0	1.17E-16	0.0065	0.0003	0.0002	0
	DBI-m	1.0867	1.0951	1.4513	1.4873	**0.74337**	1.2232
	DBI-s	2.34E-16	2.34E-16	0.1386	0	0.0046	0.0003

Note: *-m and *-s denote the values of the mean and the standard deviation, respectively.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE IX

Correspondences Between FM-CDDM, NMI, and RI on Dataset TD2 (in the Ascending Order of the Means of FM-CDDM)

FM-CDDM-mean	0.00271	0.00348	0.00426	0.00503	0.00581	0.00777
FM-CDDM-std	5.753E-12	3.283E-12	6.455E-12	9.554E-13	2.418E-12	7.090E-13
NMI-mean	0.64061	0.63896	0.63837	0.63816	0.63927	0.63860
NMI-std	0	0	0	0	0	0
RI-mean	0.90631	0.90601	0.90589	0.90583	0.90603	0.90411
RI-std	0	0	0	0	1.124E-16	0

IEEE Trans Cybern. Author manuscript; available in PMC 2017 January 01.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE X

Correspondences Between FM-CDDM, NMI, and RI on Dataset ORL (in the Ascending Order of the Means of FM-CDDM)

FM-CDDM-mean	1.761E-05	2.888E-04	3.936E-04	3.916E-04	4.780E-04	5.474E-04
FM-CDDM-std	3.861E-13	1.182E-11	1.501E-11	1.538E-11	8.326E-12	4.929E-12
NMI-mean	0.50611	0.50611	0.50611	0.50611	0.52054	0.52054
NMI-std	0	0	0	0	0	0
RI-mean	0.76596	0.76596	0.76596	0.76596	0.77748	0.77748
RI-std	0	0	0	0	0	0

IEEE Trans Cybern. Author manuscript; available in PMC 2017 January 01.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE XI

Correspondences Between FM-CDDM, NMI, and RI on Dataset RTHM (in the Ascending Order of the Means of FM-CDDM)

FM-CDDM-mean	1.031E-2	1.144E-2	1.215E-2	1.344E-2	1.393E-2	1.399E-2
FM-CDDM-std	5.759E-12	4.534E-12	6.040E-12	1.483E-11	5.373E-09	2.462E-08
NMI-mean	0.59089	0.57838	0.65040	0.62803	0.64682	0.63962
NMI-std	0	0	0.00126	0.01648	0.00182	0.00157
RI-mean	0.85854	0.85442	0.89190	0.87966	0.88904	0.88724
RI-std	0	0	0.00613	0.01111	0.00251	0.01293