Synergetic information bottleneck for joint multi-view and ensemble clustering

Xiaoqiang Yan[a], Yangdong Ye[a,*], Xueying Qiu[a], Hui Yu[b]

[a] *School of Information Engineering, Zhengzhou University, Zhengzhou, 450052, China*
[b] *School of Creative Technologies, University of Portsmouth, PO1 2DJ, United Kingdom*

Abstract

Multi-view and ensemble clustering methods have been receiving considerable attention in exploiting multiple features of data. However, both of these methods have their own set of limitations. Specifically, the performance of multi-view clustering may degrade due to the conflict between heterogeneous features, while ensemble clustering relies heavily on the quality of basic clusterings since it discovers the final clustering partition without considering the original feature structures of the source data. In this study, we propose a novel clustering scheme called synergetic information bottleneck (SIB) for joint multi-view and ensemble clustering. First, the proposed SIB utilizes multiple original features to characterize data information from different views while exploiting the basic clusterings to relieve the conflict of heterogeneous features. Second, the SIB generally formulates the problem of joint multi-view and ensemble clustering as a function of mutual information maximization, in which the relatedness between the original features and auxiliary basic clusterings is maximally preserved with respect to the final clustering partition. Finally, to optimize the objective function of SIB, a novel "draw-and-merge" optimization method is proposed. In addition, we prove that this novel optimization method can ensure that the objective function of SIB converges to a stable optimal in a finite number of iterations. Extensive experiments conducted on several practical tasks demonstrate that the SIB outperforms the state-of-the-art multi-view and ensemble clustering methods.

Keywords: Multi-view clustering, ensemble clustering, information bottleneck, mutual information

1. Introduction

In various applications, data records are usually represented by various feature descriptors [1]. For instance, a piece of news can be translated into multiple languages; objects in images can be characterized by shape, colour and texture features; and human actions in videos can usually be clarified using oriented gradient, optical flow and motion boundary features. Each singular feature can describe the data information from a certain viewpoint. However, the use of a singular feature does not lead to consistently satisfactory performance on all practical tasks due to the biases of each feature. Therefore, it has become popular to develop learning algorithms to automatically integrate multiple features.

In recent times, multi-view clustering (MVC) and ensemble clustering (EC) have been receiving considerable attention in exploiting multiple features of data. Both methods intend to improve the quality of the final clustering partition by integrating the multiple features of the same input data. Specifically, MVC methods [2, 3, 4, 5, 6, 7, 8] aim to group the data into different categories based on the relevant information from multiple features, wherein each feature can be regarded as one "view" to observe an data object. It is noteworthy that the multi-view in this study indicates multiple features rather than multiple modalities. To avoid ambiguity, we define the terms multi-view as follows:

Definition 1. *Suppose there exists a data collection X and its multiple features $F_1, F_2, \cdots, F_N$. We define the term multi-view to pertain to $F_1, F_2, \cdots, F_N$ since these features can describe N different aspects of the data characteristics.*

[*] Corresponding author.
Email addresses: `iexqyan@gmail.com` (Xiaoqiang Yan), `ieydye@zzu.edu.cn, 0086-138-3838-2185` (Yangdong Ye), `iexyqiu@gs.zzu.edu.com` (Xueying Qiu), `hui.yu@port.ac.uk` (Hui Yu)

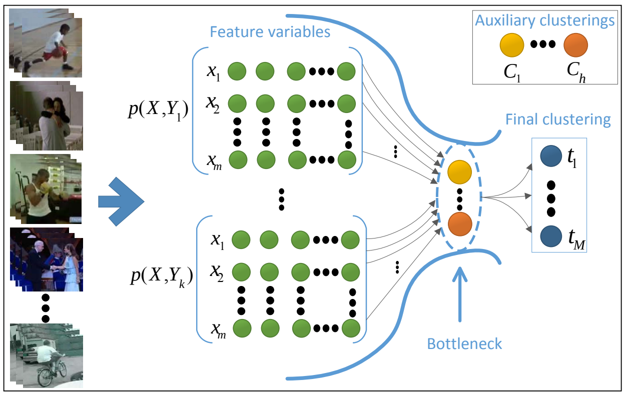


Figure 1: An illustration of the SIB method. Given an unlabelled data collection $\{x_1, \cdots, x_m\}$, SIB learns the pattern categories $\{t_1, \cdots, t_M\}$ from multiple feature variables $Y_1, \ldots, Y_k$ while considering various auxiliary clustering $\{C_1, \cdots, C_h\}$. The complementary information between multiple feature variables and basic clusterings can be preserved via the information bottleneck. (Best viewed in colour)

The existing MVC approaches can be divided into two categories, namely, function-based and subspace-based methods. The function-based methods [3, 4, 5, 7, 8] make use of certain global objective functions to optimize the cluster structure shared by multiple views. In contrast, subspace-based methods [2, 6] first project multi-view data into a common lower dimensional subspace and later exploit certain clustering techniques to discover the optimal data partition based on the common subspace. However, both types of MVC methods directly take multiple features of the raw data as input views. It is difficult to bridge the distributional gap among multiple heterogeneous features by simply combining them. For instance, scale invariant feature transform (SIFT), three patch local binary patterns (TPLBP) and colour attention (CA) features are heterogeneous to each other [9, 10], since they describe shape, texture and colour characteristics of images. In addition, the feature representations of data are often described using high-dimensional vectors, and the problem of the dimensionality curse needs to be overcome when dealing with multiple features simultaneously.

Ensemble clustering (EC) [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21], also known as consensus clustering, intends to combine multiple basic clusterings into a potentially better and more robust clustering partition, where each set of basic clusterings can be generated using one type of feature representation. EC methods integrate multiple basic clusterings into a consensus one without accessing the raw data features. Thus, this approach inevitably has the ability of leveraging complementary information from heterogeneous features [18]. In general, EC methods can be roughly divided into two categories, i.e., those with implicit objectives and explicit objectives. The first category involves directly adopting some heuristics to determine approximate solutions, such as the graph model [11, 12] and the co-association matrix [13, 17]. The other category involves employing a utility function to measure the similarity among basic partitions and the consensus one, e.g., non-negative matrix factorization [14] and k-means-like algorithm [18]. However, ensemble clustering yields the final clustering according to only the known partitions, while ignoring the original feature structures of the data. In other words, the quality of the final partition relies heavily on the basic clusterings, which means that the performance of the ensemble clustering is always sensitive to the basic clusterings.

In this study, we propose a novel clustering scheme called the synergetic information bottleneck (SIB) for joint multi-view and ensemble clustering. SIB aims to find the final clustering partition by considering multiple original features and auxiliary basic clusterings simultaneously. The original features characterize the data information from different views, while the basic clusterings clarify the data information from heterogeneous features. Specifically, SIB generally formulates the considered problem as a function of mutual information maximization. In this objective function, the information filtered from the original features and auxiliary basic clusterings is maximally preserved through a "bottleneck" with respect to the final clustering partition (see Fig. 1). Additionally, to optimize the objective function of the SIB algorithm, a novel draw-and-merge optimization method is proposed, which ensures the convergence of the objective function of the SIB. The results of performed experiments demonstrate the effectiveness of the SIB algorithm when applied to the tasks of publication and multilingual corpus analysis, object category discovery in images and unsupervised human action categorization in videos. The main contributions of this study correspond to the following four aspects:

- A novel clustering scheme called the synergetic information bottleneck (SIB) is proposed, which can cope with the overreliance of ensemble clustering on basic clusterings and mitigate the conflict between heterogeneous features in multi-view clustering.
- An extensional measurement based on mutual information is proposed to capture the relatedness between multiple original feature variables and auxiliary basic clusterings, which is a general approach and can be beneficial to many other related fields, such as cross-domain adaptation, transfer learning, multi-task learning and alternative cluster analysis.
- In the SIB, to realize the optimization of mutual information maximization, a novel sequential "draw-and-merge" solution is proposed to update the data partition. Further, we prove that this novel optimization method can ensure that the objective function of the SIB converges to a stable optimal in a finite number of iterations.
- The results of experiments performed involving several challenging tasks, including publication and multilingual corpus analysis, object category discovery in images and unsupervised human action categorization in videos, demonstrate that the proposed SIB achieves better results than the existing state-of-the-art clustering methods.

The remainder of this paper is organized as follows. In Section 2, the related work is introduced. In Section 3, we elaborate the formulation of the SIB in detail. The optimization procedure is described in Section 4. The experimental settings and comparison results pertaining to several practical tasks are reported and analysed in Section 5. Finally, Section 6 presents the conclusions of the study.

2. Related Work

2.1. Multi-view Clustering

We briefly introduce the related multi-view clustering methods [22] in terms of integrating information from multiple views. The early approaches involved integrating multiple information from different views by constructing a similarity matrix among them. For instance, Cai et al. [9] constructed a graph Laplacian matrix to integrate different models by treating each view as one model. Wang et al. [5] proposed that a universal feature embedding of all views should be generated, with the unary embedding cost and pairwise disagreement cost minimized using minimax optimization. Furthermore, co-training is an interesting approach for realizing multi-view clustering. Kumar et al. [3, 4] extended co-training and co-regularization to a multi-view clustering scenario, which could search for clusterings that are consistent across views. Another type of MVC method is the subspace-based approach, which assumes that the input views are generated from a latent view and attempt to obtain a latent subspace shared by multiple views. For example, Chaudhuri et al. [2] utilized the canonical correlation analysis (CCA) to project the data in each view to a lower-dimensional subspace. Cao et al. [6] extended subspace clustering into the multi-view domain and utilized the Hilbert-Schmidt independence criterion (HSIC) as a diversity term to explore the complementarity of multi-view representations. Some other multi-view clustering approaches, such as minimax optimization [5], belief propagation [7] and kernel spectral clustering [8], also obtained promising results. However, the performance of MVC has always been limited by the conflict between heterogeneous features.

2.2. Ensemble Clustering

In the past decades, a variety of ensemble clustering methods have been proposed. Strehl et al. [11] developed three graph-based algorithms, which solved the cluster ensemble by defining a mutual information-based objective function that enabled automatic selection of the best solution. Following this approach, Fern et al. [12] developed a bipartite graph to improve the clustering quality. Another class of ensemble clustering is based on the similarity matrix. For instance, Fred et al. [13] explored the concept of evidence accumulation clustering, which summarized the information of basic clusterings into a co-association matrix. Zhou et al. [19] constructed a connective matrix of any two instances belonging to the same class in which the Kullback-Leibler divergence was utilized as the consensus measurement. It is worth mentioning that tremendous research efforts have been devoted to constructing global objective functions for ensemble clustering. For example, Wu et al. [18] established a general framework of k-means-based ensemble clustering in which the utility function was derived from a continuously differentiable convex function. In addition, other interesting researches exist for ensemble clustering, such as the link-based cluster ensemble [15] and Bayesian ensemble [16]. However, all the ensemble clustering methods mentioned above generate the final clustering according to only the known partitions without considering the original feature structures of the data, which makes the performance of the ensemble clustering sensitive to the basic clusterings.

Recently, several studies have focused on solving MVC by employing an ensemble clustering approach [23, 24, 25, 26] in which first the basic clusterings for each view are generated individually by multi-view clustering algorithms, and next, a consensus partition among all the basic clusterings is built by ensemble clustering algorithms. Thus, these ensemble clustering approaches for solving MVC are also limited to the quality of basic clusterings. In this study, rather than utilizing ensemble clustering to solve multi-view data analysis, we focus on performing MVC and EC simultaneously under the constraint of the information bottleneck framework in this study, which maximally preserves the information filtered from the original features and auxiliary basic clusterings through a "bottleneck" with respect to the final clustering partition.

2.3. Information Bottleneck

The information bottleneck (IB) [27] is an unsupervised model independent data organization technique. To make this paper more self-contained, we summarize it from the perspective of its multivariate extension. Given a set of random variables $\mathbf{X} = \{X_1, \cdots, X_n\}$, the multivariate IB [28] aims to find a set of partitions $\mathbf{T} = \{T_1, \cdots, T_k\}$ by searching for the distribution $q(\mathbf{T}|\mathbf{X})$, where $\mathbf{X}$ is compressed to $\mathbf{T}$ as much as possible. The multivariate IB method utilizes two *Bayesian networks* with graph G_{in} and G_{out} to denote the systems of clusters and what information should be maintained. The graph G_{in} is defined over $\mathbf{X} \cup \mathbf{T}$, and it defines a distribution $q(\mathbf{X}, \mathbf{T}) = q(\mathbf{T}|\mathbf{X})p(\mathbf{X})$, which specifies the compression relationships between $\mathbf{X}$ and $\mathbf{T}$. The other graph G_{out} is also defined over $\mathbf{X} \cup \mathbf{T}$, and it specifies the relevant information that $\mathbf{T}$ is expected to be able to preserve. The multivariate IB utilizes *multi-information* to calculate the information between multiple variables in these two networks. The multi-information of $\mathbf{X}$ is defined as follows:

$$\mathcal{I}(\mathbf{X}) = D_{KL}[p(X_1, \cdots, X_n) \| p(X_1), \cdots, p(X_n)], \tag{1}$$

where D_{KL} is the Kullback-Leibler divergence [29]. The multivariate IB is suggested to minimize the function

$$\mathcal{L}_{min}(q(\mathbf{T}|\mathbf{X})) = \mathcal{I}^{G_{in}}(\mathbf{X}, \mathbf{T}) - \beta \cdot \mathcal{I}^{G_{out}}(\mathbf{X}, \mathbf{T}), \tag{2}$$

where $\mathcal{I}^G$ is the multi-information concerning a Bayesian network structure G over $\mathbf{X} \sim p(\mathbf{X})$, which is defined as follows:

$$\mathcal{I}^G(\mathbf{X}) = \sum_i I(X_i; \mathbf{Pa}^G_{X_i}), \tag{3}$$

where $\mathbf{Pa}^G_{X_i}$ is the set of parents of X_i, and $I(X_i, \mathbf{Pa}^G_{X_i})$ is the mutual information between X_i and its parents $\mathbf{Pa}^G_{X_i}$.

The relationship between data compression and information preservation in the original IB is shown in Fig. 2. We can obtain $\mathcal{I}^{G_{in}} = I(X; T) + I(X; Y)$ and $\mathcal{I}^{G_{out}} = I(T; X) + I(T; Y)$. IB has demonstrated its superiority in many multivariate problems such as multi-view learning [1], multi-feature analysis [30] and multi-task learning [31]. Recently, IB theory has been applied to the ensemble clustering method [32]. However, [32] focused only on the ensemble setting, while the proposed SIB adopts a new objective function and optimization method for the task of joint multi-view and ensemble clustering. In addition, several researchers have attempted [33, 34, 35] to deal with multiple variables by using the IB method. Since the IB originates from the rate-distortion theory, it is natural to treat multiple variables as different information sources for the input of the IB method. To the best of our knowledge, the present study is the first work to integrate multi-view and ensemble clustering together under the constraint of the IB principle.

3. Synergetic Information Bottleneck for Joint Multi-view and Ensemble Clustering

In this section, we elaborate on the proposed synergetic information bottleneck (SIB) method. First, we define the problem of joint multi-view and ensemble clustering via the SIB method. Next, the objective function of the SIB is presented in detail.

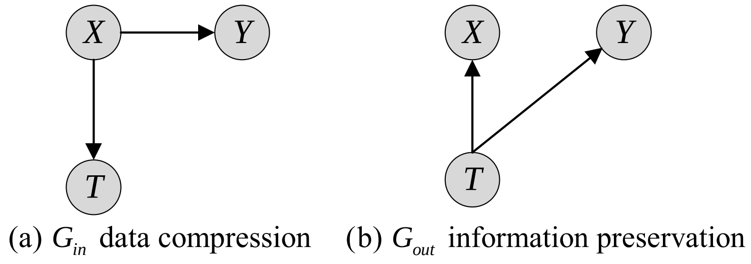

X
Y
T
(a) G_{in} data compression
X
Y
T
(b) G_{out} information preservation

Figure 2: **Information bottleneck** method. (a) The Bayesian network G_{in} expresses the compression from X to T, while Y characterizes the relevant information concerning X. (b) The Bayesian network G_{out} specifies the systems of clusters T and the information terms corresponding to the feature variable Y that should be maintained.

3.1. Problem Formulation

Let $X = \{x_1, x_2, \ldots, x_m\}$ denote an unlabeled data set that can be partitioned into different categories or clusters, where m is the number of data elements. For the data representation, various types of feature descriptors are available to depict different aspects of the source data, such as multiple translations in multilingual corpus; local shape, colour, texture information in images; and oriented gradient, optical flow, and motion boundary features in videos. We use discrete random variables $F_1, \ldots, F_{k+h}$ $(k, h \geq 1)$ to denote $k + h$ types of features. First, the $k + h$ feature variables are divided into two parts. One part with k original feature variables $Y_1, \ldots, Y_k$ is utilized to characterize the data information from different views, while the remaining h feature variables are adopted to generate multiple basic clusterings $C_1, \ldots, C_h$ of the source data. In particular, the i-th original feature variable Y_i takes values from one feature source $Y_i = \{y^i_1, y^i_2, \ldots, y^i_d\}$ $(1 \leq i \leq k)$, which characterizes the source data X from the i-th view. The j-th clustering C_j $(1 \leq i \leq h)$ is the clustering (cluster assignment) constructed using the j-th features of the remaining h feature variables. Thus, the goal of the SIB method is to learn an assignment $q(T|X)$ from X to its cluster partition T by considering the multiple original features $Y_1, \ldots, Y_k$ and auxiliary basic clusterings $C_1, \ldots, C_h$ simultaneously.

3.2. Objective Function of SIB

In this section, we define the objective function of the proposed SIB, which can learn the cluster structures hidden in source data by performing multi-view and ensemble clustering simultaneously. The SIB treats the clustering procedure as a process of data compression. Given the original feature variables $Y_1, \ldots, Y_k$ and basic clusterings $C_1, \ldots, C_h$, SIB aims to compress the source random variable X to its compressed representation T as much as possible, while the compressed variable T ought to maximally preserve the relevant information with the original feature variables and auxiliary basic clusterings. Here, we utilize two Bayesian networks G_{in} and G_{out} to denote the systems of clusters and the relevant information that should be preserved. As shown in Fig. 3, the SIB model consists of two graphs: G_{in} data compression and G_{out} information preservation. In the first graph, the source data collection X is required to be compressed into its compressed variable T. In the second graph, the compressed variable T ought to maximally preserve the information concerning the multiple original feature variables and basic clusterings. The amount of information in data compression and information preservation can be defined as follows:

$$\begin{cases} \mathcal{I}^{G_{in}} = I(T; X) \\ \mathcal{I}^{G_{out}} = \sum_{i=1}^{k} I(T; Y_i) + \sum_{j=1}^{h} I(T; C_j). \end{cases} \tag{4}$$

According to function (2) of the multivariate IB, the SIB algorithm can be generally formulated as follows:

$$\begin{aligned} &\mathcal{L}_{min}[q(T|X)] = \mathcal{I}^{G_{in}} - \beta \cdot \mathcal{I}^{G_{out}} = \\ &I(X; T) - \beta \cdot [\sum_{i=1}^{k} I(T; Y_i) + \sum_{j=1}^{h} I(T; C_j)], \end{aligned} \tag{5}$$

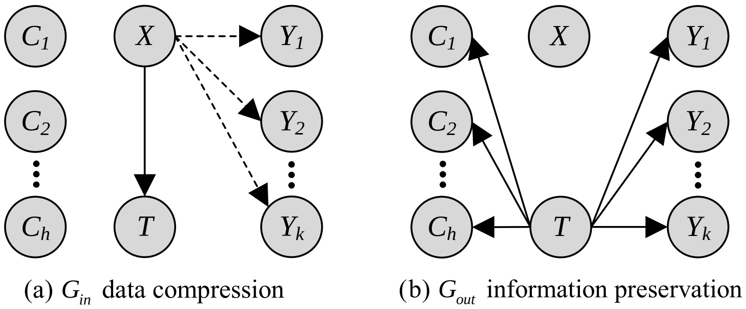


Figure 3: **Synergetic Information Bottleneck** method. (a) The Bayasian network G_{in} shows the compression from source variable X to its compressed representation T. $Y_1, \ldots, Y_k$ denotes the multiple relevant feature variables. $C_1, \ldots, C_h$ are multiple auxiliary basic clusterings constructed by the remaining h original features. (b) The Baysian network G_{out} implies that the compressed variable T should preserve the information with respect to original feature variables $Y_1, \ldots, Y_k$ and auxiliary basic clusterings $C_1, \ldots, C_h$.

where $I(X;T)$ is the mutual information measuring the compactness of X to its compressed representation T. The notion $\sum_{i=1}^{k} I(T;Y_i)$ measures the amount of information that variable T maintains about the original features, while $\sum_{j=1}^{h} I(T;C_j)$ measures the information concerning the basic clusterings $C_1, \ldots, C_h$ that is contained in variable T. β strikes a balance between the compression in G_{in} and the relevant information preservation in G_{out}. For the convenience of optimization, we divide both sides of Equation (5) by $-\beta$ and obtain the following objective function

$$\begin{aligned}&\mathcal{L}_{max}[q(T|X)] = \mathcal{I}^{G_{out}} - \beta^{-1} \cdot \mathcal{I}^{G_{in}} = \\ &\sum_{i=1}^{k} I(T;Y_i) + \sum_{j=1}^{h} I(T;C_j) - \beta^{-1} \cdot I(X;T).\end{aligned} \tag{6}$$

It can be observed that the remaining task of the SIB is to maximize the objective function (6). To ensure the convergence of this function, we present a sequential information-theoretic solution, which always performs better than agglomerative methods do. In this study, we consider the hard clustering manner, which means the value of $q(T|X)$ is either 0 or 1.

4. Optimization Method

In this section, we propose a novel draw-and-merge optimization method to solve the objective function of the SIB. First, the draw-and-merge procedure, which ensures that the SIB algorithm converges into an optimal solution, is presented. Next, the relatedness measurement between the original feature variables and the existing basic clusterings is given. Finally, the algorithm and its relevant analyses, such as convergence and complexity, are introduced.

4.1. Draw-and-Merge Procedure

To solve the problem of maximizing objective function (6), we propose a sequential information-theoretic optimization, which is essentially the "draw-and-merge" procedure. The draw-and-merge method first stochastically partitions the source variable X into u clusters. At the following iterative step, each $x \in X$ is drawn from its original category t^{old} and treated as a new cluster $\{x\}$. Thus, the number of clusters becomes $u + 1$. The singleton cluster $\{x\}$ should be merged into t^{new} to realize the information loss minimization. In other words, the sequential information-theoretic optimization method must increase the value of function (6) of the SIB algorithm.

In the draw-and-merge procedure, we attempt to merge the singleton cluster $\{x\}$ into an optimal cluster t^{new} at each step. For clarity, let $\mathcal{L}^{bef}$ and $\mathcal{L}^{aft}$ respectively denote the value of function (6) before and after the single x is drawn from its original category. Let $\mathcal{L}^{new}$ be the value of function (6) after the single x is merged into a certain cluster t^{new}. In the merge step, selecting an optimal cluster t^{new} means choosing the minimum value change between $\mathcal{L}^{aft}$ and

$\mathcal{L}^{new}$. Here, the value change of the objective function (6) in one draw-and-merge procedure is called "merger cost", i.e., $d_{\mathcal{L}} = \mathcal{L}^{aft} - \mathcal{L}^{new}$, which is defined as follows according to function (6):

$$\begin{aligned}
d_{\mathcal{L}} &= \mathcal{L}^{aft} - \mathcal{L}^{new} \\
&= \sum_{i=1}^{k}[I(T^{aft};Y_i) - I(T^{new};Y_i)] + \sum_{j=1}^{h}[I(T^{aft};C_j) - I(T^{new};C_j)] - \beta^{-1}\cdot[I(T^{aft};X) - I(T^{new};X)] \\
&= \sum_{i=1}^{k}\Delta I^{i}_{view} + \sum_{j=1}^{h}\Delta I^{j}_{clustering} - \beta^{-1}\cdot \Delta I_{com}
\end{aligned} \tag{7}$$

To maximize the value of objective function (6), we merge $\{x\}$ into t^{new} such that $t^{new} = \arg\min d_{\mathcal{L}}$. Next, we define the following proposition:

Proposition 1. *Let $\{x\}$ be a singleton cluster, t be the cluster that $\{x\}$ will be merged into and t^{new} be the new cluster after the merging , i.e., $\{\{x\},t\} \Rightarrow t^{new}$; we have*

$$\begin{aligned}
p(t^{new}) &= p(x) + p(t) \\
p(Y_i|t^{new}) &= \frac{p(x)}{p(t^{new})}p(Y_i|x) + \frac{p(t)}{p(t^{new})}p(Y_i|t),
\end{aligned} \tag{8}$$

where $1 \leq i \leq k$.

Now, let ΔI_{view} be the value change in objective function (6) caused by the $I(Y_i;T)$ term. Suppose $\{x\}$ is merged into the cluster t to generate a new cluster $\widetilde{t}$, i.e. $\{\{x\},t\} \Rightarrow \widetilde{t}$,

$$\begin{aligned}
\Delta I^{i}_{view} &= I(T^{aft};Y_i) - I(T^{new};Y_i) \\
&= p(t)\sum_{y} p(y_i|t)\log\frac{p(y_i|t)}{p(y_i)} + p(x)\sum_{y_i} p(y_i|x)\log\frac{p(y_i|x)}{p(y_i)} \\
&- p(\widetilde{t})\sum_{y_i} p(y_i|t^{new})\log\frac{p(y_i|t^{new})}{p(y_i)}.
\end{aligned}$$

Using Proposition 1, the value change ΔI^{i}_{view} can be derived as follows:

$$\begin{aligned}
\Delta I^{i}_{view} &= p(t)\sum_{y_i} p(y_i|t)\log\frac{p(y_i|t)}{p(y_i)} + \\
& p(x)\sum_{y} p(y_i|x)\log\frac{p(y_i|x)}{p(y_i)} - \sum_{y_i} p(t)p(y_i|t)\log\frac{p(y_i|t^{new})}{p(y_i)} - \\
& \sum_{y_i} p(x)p(y_i|x)\log\frac{p(y_i|t^{new})}{p(y_i)} \\
&= p(t)\sum_{y_i} p(y_i|t)\log\frac{p(y_i|t)}{p(y_i|t^{new})} + \\
& p(x)\sum_{y_i} p(y_i|x)\log\frac{p(y_i|x)}{p(y_i|t^{new})} \\
&= p(t)D_{KL}[p(y_i|t)||p(y_i|t^{new})] \\
&+ p(x)D_{KL}[p(y_i|x)||p(y_i|t^{new})] \\
&= [p(t) + p(x)]JS_{\prod}[p(y_i|t), p(y_i|x)],
\end{aligned}$$

where $JS_{\prod}$ is the $Jensen\text{-}Shannon$ (JS) divergence [29] used to measure the similarity between two probability distributions, $\prod = \{\frac{p(x)}{p(x)+p(t)}, \frac{p(t)}{p(x)+p(t)}\}$. We can consider that $\Delta I^{i}_{view} \geq 0$ since $JS_{\prod} \geq 0$.

Algorithm 1 SIB Algorithm

```
Input: Joint distributions $p(X, Y_1), \cdots, p(X, Y_k)$; basic clusterings $C_1, \ldots, C_h$; balance parameters $\beta$; number of clusters $u$.
Output: Final clustering partition $T$ of $X$.
Initialize: Divide $X$ into $M$ clusters stochastically;
repeat
  for all $x \in X$ do
    Draw: Draw $x$ from its original category $t(x)$ and treat it as a singleton cluster $\{x\}$;
    Merger cost calculation:
    For singleton cluster $\{x\}$, compute the merger costs $d_{\mathcal{L}}$ according to Equation (7);
    Merge: Merge the single $x$ into an optimal cluster $t^{new}$ such that $t^{new} = \arg\min_{t \in \mathcal{T}} d_{\mathcal{L}}$;
  end for
until Convergence
```

Similarly, we can obtain the value change ΔI_{com} in objective function (6) caused by the $I(X;T)$ term as follows:

$$\begin{aligned} &\Delta I_{com} = I(T^{aft}; X) - I(T^{new}; X) = \\ &[p(x) + p(t)] JS_{\Pi}[p(x), p(x|t)]. \end{aligned} \tag{9}$$

In this study, we employ mutual information to measure the relatedness between feature variables and basic clusterings. Let C_i be the i-th existing basic clustering. The value change caused by $I(T; C_j)$ in function (6) can be calculated as follows:

$$\Delta I^{j}_{clustering} = I(T^{aft}; C_j) - I(T^{new}; C_j). \tag{10}$$

In the next section, we explain the detailed measurement of the relatedness between the multiple feature variables and various auxiliary basic clusterings.

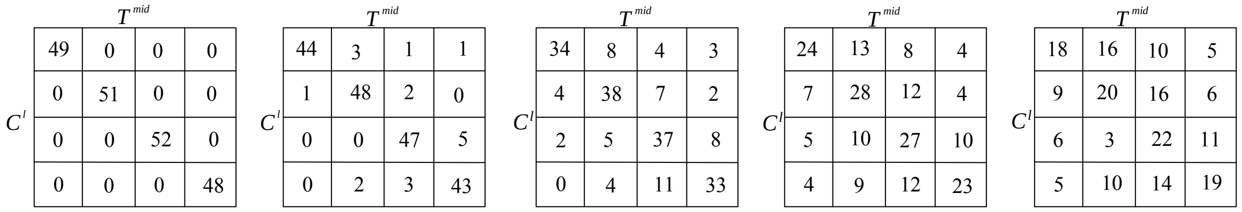
4.2. Relatedness Measurement

In the SIB framework, we intend to find the hidden cluster structure that remains in the source data by incorporating multiple feature variables and various complementary clusterings. Therefore, the relatedness measurement between the feature variables and basic clusterings is one key issue in SIB. The mutual information is an effective measurement to quantify how much "information" is contained in a variable about another one, and its effectiveness has been verified in various tasks, such as dual spectral clustering [36] and non-redundant clustering [37]. However, due to the heterogeneous structure of the feature variables and basic clusterings, the mutual information cannot be applied directly. We next describe the extensional calculation in detail.

In SIB, $k + h$ discrete random variables $F_1, \ldots, F_{k+h}$ are available to represent $k + h$ types of features of the source data. We divide the $k + h$ discrete random variables into two parts: One part with k variables is treated as feature variables $Y_1, \ldots, Y_k$; next, the remaining h features are utilized to construct multiple basic clusterings $C_1, \ldots, C_h$ of the source data. In the SIB framework, the draw-and-merge optimization is proposed to learn an optimal representation T from both the original feature variables and basic clusterings, which is an iterative procedure. In the SIB iteration involving feature variables, we use $T^{mid} = \{t_1^{mid}, t_2^{mid}, \cdots, t_u^{mid}\}$ to represent the temporary partition, where u is the number of clusters. Similarly, let C^l be one partition of multiple auxiliary clusterings $C_1, \ldots, C_h$, taking values from $C^l = \{c_1^l, c_2^l, \cdots, c_u^l\}$. To measure the relationship between the feature variable and auxiliary clustering, first, the co-occurrence matrix of the feature variables and auxiliary clusterings should be constructed.

As mentioned earlier, there are m data elements in the unlabelled data collection X, which take values from $\{x_1, x_2, \ldots, x_m\}$. Let m_i be the number of data points that are allocated into cluster t_i^{mid}, let m_j be the number of data points that are allocated into cluster c_j^l, and let m_{ij} be the number of data points that are allocated into clusters t_i^{mid} and c_j^l at the same time. The joint co-occurrence distribution of cluster T^{mid} and C^l can be computed as follows:

$$\begin{cases} p(t_i^{mid}) = m_i/m, \\ p(c_j^l) = m_j/m, \\ p(t_i^{mid}, c_j^l) = m_{ij}/m. \end{cases} \tag{11}$$



49	0	0	0
0	51	0	0
0	0	52	0
0	0	0	48

(a). $I(T^{mid};C^l)=1.3858$

44	3	1	1
1	48	2	0
0	0	47	5
0	2	3	43

(b). $I(T^{mid};C^l)=1.0037$

34	8	4	3
4	38	7	2
2	5	37	8
0	4	11	33

(c). $I(T^{mid};C^l)=0.5182$

24	13	8	4
7	28	12	4
5	10	27	10
4	9	12	23

(d). $I(T^{mid};C^l)=0.1777$

18	16	10	5
9	20	16	6
6	3	22	11
5	10	14	19

(e). $I(T^{mid};C^l)=0.0764$

Figure 4: Co-occurrence matrix and corresponding mutual information between feature variables and auxiliary clusterings.

Given multiple feature variables and complementary basic clusterings, we can calculate their mutual information as follows:

$$I(T^{mid};C^l) = \sum_{t_i^{mid}\in T^{mid}} \sum_{c_j^l\in C^l} p(t_i^{mid},c_j^l)\log\frac{p(t_i^{mid},c_j^l)}{p(t_i^{mid})p(c_j^l)}. \tag{12}$$

In Fig. 4, we present an example to demonstrate the effectiveness of the variable measurement. As shown in Fig. 4 (a), the mutual information between the feature variable and clustering variable is relatively high, because the partition according to the feature variable is the same as that for the complementary clustering. With an increase in the difference between feature variable and auxiliary clustering, the mutual information gradually declines. Thus, mutual information can effectively measure the difference between feature variables and basic clusterings. The pseudo-code of the SIB can be described, as shown in Algorithm 1.

4.3. Theoretical Analysis

In this section, we first prove that the objective function of SIB can converge to a stable solution in a finite number of iterations and later derive the computation costs of the SIB.

Theorem 1. *In the SIB framework, let $\mathcal{L}(x,t^{bef})$ be the value of the SIB objective function before x is drawn from its original cluster t^{bef}, and let $\mathcal{L}(x,t^{new})$ denote the value of the SIB objective function after the singleton cluster $\{x\}$ is merged into cluster t^{new}. The following expression can be obtained:*

$$\mathcal{L}(x,t^{new}) \geq \mathcal{L}(x,t^{bef}). \tag{13}$$

Proof. Let $\mathcal{L}(x,t^{mid})$ be the value of the SIB objective function after x is drawn from some clusters t^{bef}. In the draw-and-merge procedure, x must be merged into a certain cluster t^{new} such that $t^{new} = \arg\min d_{\mathcal{L}}(\{x\},t^{new})$, where $d_{\mathcal{L}}$ pertains to the information loss. If $t^{bef} = t^{new}$, it is implied that x is merged into the original cluster t^{bef}, and thus, $\mathcal{L}(x,t^{new}) = \mathcal{L}(x,t^{bef})$. If $t^{bef} \neq t^{new}$,

$$\begin{cases} \mathcal{L}(x,t^{bef}) = \mathcal{L}(x,t^{mid}) - d_{\mathcal{L}}(\{x\},t^{bef}), \\ \mathcal{L}(x,t^{new}) = \mathcal{L}(x,t^{mid}) - d_{\mathcal{L}}(\{x\},t^{new}). \end{cases} \tag{14}$$

Note that in each merge procedure, the singleton cluster $\{x\}$ is merged into t^{new} such that $t^{new} = \arg\min d_{\mathcal{L}}(\{x\},t^{new})$, and thus, $d_{\mathcal{L}}(\{x\},t^{new}) < d_{\mathcal{L}}(\{x\},t^{bef})$. We obtain $\mathcal{L}(x,t^{new}) \geq \mathcal{L}(x,t^{bef})$. □

Corollary 1. *The objective function of the SIB algorithm can converge to a stable solution.*

Proof. According to Theorem 1, $\mathcal{L}(x,t^{new}) \geq \mathcal{L}(x,t^{bef})$, which means that all draw-and-merge procedures do not decrease the value of function (6). The general idea of the convergence proof is to show that the objective function (6) of the SIB to be upper-bounded. Assuming that the source data X has a clustering partition C and original feature Y, we obtain the following equation:

$$\begin{aligned}\mathcal{L}_{max} &= \sum_{i=1}^{k} I(T;Y_i) + \sum_{j=1}^{h} I(T;C_j) - \beta^{-1} \cdot I(X;T) \\ &\approx k \cdot I(T;Y) + h \cdot I(T;C) - \beta^{-1} I(X;T) \\ &= -\frac{1}{\beta}[I(X;T) - k\beta I(T;Y)] + h \cdot I(T;C)\end{aligned} \tag{15}$$

For a given k and β, the term $I(T;X) - k\beta I(T;Y)$ is equivalent to the objective function of the original IB, i.e., $I(X;T) - \beta \cdot I(T;Y)$, which has been proven to be lower-bounded in [38]. Dividing the equation by $-\beta$, the first term $-\frac{1}{\beta}[I(T;X) - k\beta I(T;Y)]$ can be seen to be upper-bounded. Assuming that C_{truth} is a ground-truth partition of the source data, we can obtain $h \cdot I(T;C) \leq h \cdot I(T;C_{truth})$, i.e., the second term $h \cdot I(T;C)$ is also upper-bounded. Thus, the objective function of the SIB algorithm can converge to a stable solution in a finite number of iterations. □

Next, we analyse the time complexity of the SIB. At step 3, the source data X are partitioned into different clusters with random initialization, thus, this step takes linear time $O(|X|)$, where $|X|$ is the number of data points in X. In the main loop the complexity of the drawing data point x at step 6 is also $O(|X|)$. The computation of the merge cost in step 7 takes time $O(u|X|(|Y_1| + \cdots + |Y_k|))$, where u is the number of clusters. The calculation of the mutual information between t^{new} and multiple clusterings $C_1, \ldots, C_h$ takes time $O(1)$. Thus, the time complexity of the SIB is $O(u|X|(|Y_1| + \cdots + |Y_k|))$.

5. Experiments

In this section, we present the experimental results pertaining to the application of the SIB method for three practical tasks, i.e., publication and multilingual corpus analysis, object category discovery in images and unsupervised human action categorization in videos. Specifically, first the clustering effectiveness and quality of SIB are demonstrated, and later, the impact of several major factors on the performance of SIB is investigated.

5.1. Experimental Settings

5.1.1. Datasets and Features

The datasets used in our experiments were extracted from three domains, that is, documents, images and videos. These datasets have diverse feature types, as presented in Table 1.

Two link-based document datasets namely, CiteSeer[1], Cora[2], and one multilingual data corpus, namely, Reuters[3], where used in the experiments. CiteSeer and Cora consist of scientific publications classified into different classes. Each publication is described by content (title and abstract) and citation. For CiteSeer, the title and abstract of the paper are described by a 3703-dimensional 0/1-valued word vector, and the citing relationships between publications are represented by a 3309-dimensional vector. For Cora, the dimensions of the content and citations are 1433 and 5429, respectively. The documents in the multilingual corpus Reuters are initially in English (EN), and the FR, GR, IT, and SP views corresponds to the words of their translations in French, German, Italian and Spanish. The multilingual documents were selected randomly, and 2000 words were selected with the k-medoids algorithm.

Three image datasets[4], from the web (Amazon), digital SLR camera (Dslr) and web camera (Webcam), were employed to demonstrate the performance of the SIB when applied to object category discovery. We adopted the following four descriptors: Dense-SIFT[5], speeded up robust features (SURF) [39], three patch local binary patterns (TPLBP) [40] and Colour Attention (Colour) [41] to represent the image collections. The popular bag-of-feature

[1] https://linqs-data.soe.ucsc.edu/public/lbc/citeseer.tgz
[2] https://linqs-data.soe.ucsc.edu/public/lbc/cora.tgz
[3] http://membres-lig.imag.fr/grimal/data.html
[4] http://www.eecs.berkeley.edu/mfritz/domainadaptation/
[5] http://www.vlfeat.org/overview/dsift.html

Table 1: Statistics of the evaluated datasets

Datasets	#Instances	#Classes	#Features	#Dimensions
CiteSeer	3312	6	2	3307, 3309
Cora	2708	7	2	1433, 5429
Reuters	1200	6	5	2000
Amazon	2813	31	4	1000
Dslr	489	31	4	1000
Webcam	795	31	4	1000
UCF Sports	150	10	4	1000
UCF 50	6676	50	4	1000
HMDB	6849	51	4	1000

(BoF) model [42] was utilized to transform the images into the co-occurrence vector of visual words. The vocabulary size in BoF was set to 1000, as reported in existing literature.

To evaluate the effectiveness of the proposed SIB algorithm for the task of unsupervised human action categorization in videos, the experiments were performed on three benchmark video datasets, namely, UCF Sports[6] [43], UCF50[7] [44] and HMDB[8] [45]. For the multiple representations, we adopt the following four descriptors: space-time interest points (STIP) [46], histogram of optical flow (HOF), histogram of oriented gradient (HOG) [47] and 3-dimensional SIFT descriptor (3DSIFT) [48]. In addition, the popular BoF model was utilized to transform the videos into a co-occurrence vector of the key visual words. The vocabulary size in BoF was set to 1000 as in the literature.

5.1.2. Baselines

We compare the proposed SIB algorithm with the following representative clustering algorithms: (1) Original information bottleneck (IB); (2) traditional clustering algorithms: *k*-means, pLSA [49], LDA [50] and NCuts [51]; (3) multi-view clustering algorithms: the co-regularized multi-view spectral clustering (CRSC) [4], co-training multi-view spectral clustering (CTSC)[9] [3], robust multi-view spectral clustering (RMSC) [52], multi-feature information bottleneck (MfIB) [30] and multi-view kernel spectral clustering (MVKSC)[10] [8]; and (4) ensemble clustering algorithms: cluster-based similarity partitioning algorithm (CSPA) [11], meta-clustering algorithm (MCLA)[11] [11], Bayesian cluster ensemble (BCE) [16], consensus information bottleneck (CIB) [32], multi-view kernel k-means clustering ensemble (MKCE) [25] and multi-view spectral clustering ensemble (MSCE)[12] [25]. For the task of object category discovery in images, we also compared the SIB with three state-of-the-art image clustering algorithms, i.e., local discriminant models and global integration (LDMGI)[13] [53], clustering-by-composition (CC)[14] [54] and ensemble projection (EP)[15] [55]. For the task of unsupervised human action categorization, three promising action clustering methods were further adopted as baselines, including dual assignment *k*-means (DAKM)[16] [36], multivariate video information bottleneck (MvIB) [56] and consensus information bottleneck (CIB) [32]. All the source code of the comparisons were provided by their original authors.

[6] http://crcv.ucf.edu/data/UCF_Sports_Action.php

[7] http://crcv.ucf.edu/data/UCF50.php

[8] http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/

[9] http://users.umiacs.umd.edu/ abhishek/papers.html

[10] https://www.esat.kuleuven.be/stadius/ADB/software.php

[11] http://strehl.com/soft.html

[12] http://users.iit.demokritos.gr/ gtzortzi/#pubsoft

[13] http://www.escience.cn/people/fpnie/papers.html

[14] http://www.wisdom.weizmann.ac.il/ alonf/code.html

[15] http://www.vision.ee.ethz.ch/ daid/

[16] http://smjdv.com/

Table 2: Comparison of the AC (%) of SIB with those of IB and typical clustering methods when applied to publication and multilingual corpus analysis

Datasets	IB		IB					IB	k-means	pLSA	LDA	NCuts	SIB
	Content	Cition	EN	FR	GR	IT	SP	Con					
CiteSeer	55.0	25.5	–	–	–	–	–	53.7	47.9	45.0	29.4	44.0	**63.7**(↑)
Cora	48.9	40.0	–	–	–	–	–	50.9	32.8	34.5	22.2	41.2	**57.8**(↑)
Reuters	–	–	50.3	52.7	52.8	53.1	50.8	55.7	22.5	41.7	53.5	43.0	**75.4**(↑)
Average	52.0	32.8	50.3	52.7	52.8	53.1	50.8	53.4	34.4	40.4	35.0	42.7	**65.6**(↑)

Table 3: Comparison of the AC (%) of SIB with those of multi-view and ensemble clustering methods when applied to publication and multilingual corpus analysis

Datasets	Multi-view clustering					Ensemble clustering							SIB
	CTSC	CRSC	RMSC	MfIB	MVKSC	CSPA	HGPA	MCLA	BCE	CIB	MKCE	MSCE	
CiteSeer	57.6	41.6	53.2	50.2	41.3	59.0	33.7	59.3	56.7	60.6	51.62	58.20	**63.7**(↑)
Cora	51.1	30.6	53.7	55.8	40.0	52.5	41.7	51.9	52.3	55.9	44.38	55.19	**57.8**(↑)
Reuters	69.0	67.2	63.2	66.0	64.8	40.6	46.7	63.5	51.3	70.6	58.33	61.83	**75.4**(↑)
Average	59.2	46.5	56.7	57.3	48.7	50.7	40.7	58.2	53.4	62.4	51.44	58.41	**65.6**(↑)

5.1.3. Evaluation Criteria

In this study, we used the normalized mutual information (NMI) [11] and clustering accuracy (AC) [57] to evaluate the clustering results. The NMI suggested in [11] is defined as follows:

$$NMI = \frac{I(L,T)}{\sqrt{H(L)H(T)}}, \tag{16}$$

where L denotes the known labels and T represents the clustering results of learning algorithms; $I(L,T)$ is the mutual information between L and T; and $H(L)$ and $H(T)$ denote the entropies of L and T, respectively.

The clustering accuracy (AC) is defined as

$$AC = \frac{\sum_{i=1}^{m} \delta(l_i, \mathrm{map}(t_i))}{m}, \tag{17}$$

where l_i is the truth label of data element x_i, t_i is the predicted label of x_i, and m is the total number of the data instances. The function map(t_i) maps each predicted label t_i to its truth label provided by the source data. The function $\delta(x,y) = 1$ when $x = y$, and $\delta(x,y) = 0$ otherwise.

5.2. Publication and Multilingual Corpus Analysis

This section describes the application of the SIB to the task of publication and multilingual corpus analysis. All the results are the mean values of AC and NMI over 10 runs. For the two link-based publication datasets, the SIB adopts the content text as the feature variable, while the citation is used to generate the basic clusterings. For the multilingual data, we selected three languages (EN, FR, and GR) as feature variables, while the remaining two languages were naturally utilized to construct the basic clusterings.

5.2.1. Comparison with Original IB

We compared the performance of the SIB with the original IB for the the task of publication and multilingual corpus analysis. From Table 2, the following observations can be made. First, the single feature is not sufficiently discriminative for different document datasets. In the case of the multilingual dataset, the IB algorithm performs differently for different translations, e.g., the AC value of the IB method when applied to the original language (EN) is 50.3%, while better results are obtained when it is performed on the IT translation (53.1%).

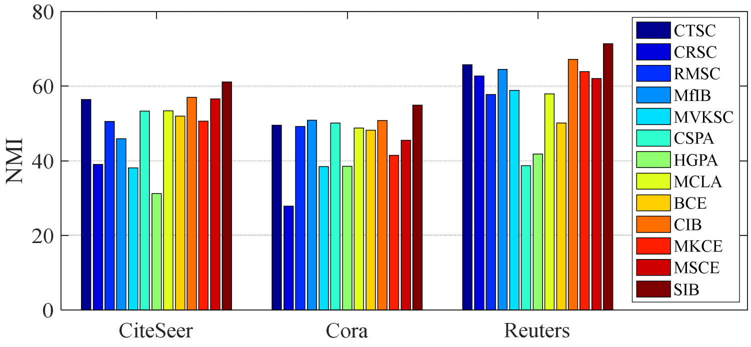


Figure 5: Comparison of the NMI (%) of SIB with those of multi-view and ensemble clustering methods when applied to publication and multilingual corpus.

Table 4: Comparison of the AC (%) of SIB with those of original IB and four other typical clustering methods when applied to object category discovery in images.

Datasets	IB				IB	*k*-means	pLSA	LDA	NCuts	SIB
	Dense-SIFT	SURF	TPLBP	Colour	Con					
Amazon	27.3	27.7	24.5	12.0	18.1	13.3	17.7	19.9	19.7	**32.9**(↑)
Dslr	43.5	43.6	18.4	34.2	44.9	32.4	33.1	31.7	31.1	**50.4**(↑)
Webcam	41.3	42.4	35.7	28.0	44.8	30.5	34.3	30.3	31.4	**49.7**(↑)
Average	37.4	37.9	26.2	24.7	35.9	25.4	28.4	27.3	27.4	**44.3**(↑)

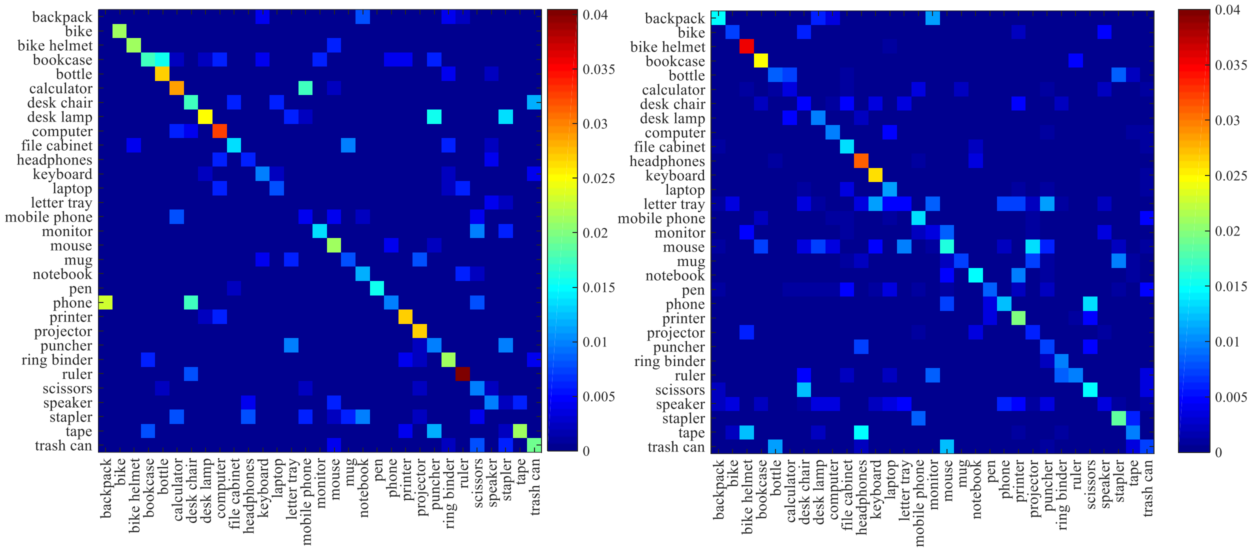
Second, when concatenating multiple features together, the original IB algorithm does not consistently demonstrate improved performances. As shown in the Con column in Table 2, the performances of the original IB algorithm when applied to the Cora and Reuters data indicate a slight improvement (2.0% and 2.6%, respectively) compared with the best AC value of the IB algorithm on the individual feature. However, for the CiteSeer data, the AC values of the IB on the combined features decrease 1.3%. Thus, by simply concatenating multiple features together, the IB method does not demonstrate consistent improved performance .

The superiority of the SIB against the original IB is also demonstrated in Table 2. Owing to the consideration of multiple original features and auxiliary basic clusterings simultaneously, the performances of the SIB algorithm are clearly better than those of the IB on all three document datasets. In particular, in comparison with the best AC values of the original IB on all single features, the SIB algorithm demonstrates improvement of 8.7% and 8.9% when applied to the two link-based publication datasets, while a significant improvement (22.3%) is attained on the multilingual document dataset.

Finally, we conducted experiments to compare the SIB with *k*-means, pLSA, LDA and NCuts algorithms, and the mean AC value of the SIB algorithm on the three document datasets exhibited considerable great improvements

Table 5: Comparison of the AC (%) of SIB with those of multi-view and ensemble clustering methods when applied to object category discovery in images.

Datasets	Multi-view clustering					Ensemble clustering							SIB
	CTSC	CRSC	RMSC	MfIB	MVKSC	CSPA	HGPA	MCLA	BCE	CIB	MKCE	MSCE	
Amazon	25.9	21.3	15.6	29.6	23.3	25.2	14.2	22.9	23.4	30.1	23.22	29.24	**32.9**(↑)
Dslr	41.7	35.4	36.1	46.9	38.9	47.8	44.7	46.0	42.4	47.8	45.04	44.32	**50.4**(↑)
Webcam	38.5	36.3	28.7	47.4	34.8	40.8	33.6	37.5	37.8	48.2	44.41	46.93	**49.7**(↑)
Average	35.4	31.0	26.8	41.3	32.3	37.9	30.8	35.4	34.5	42.0	37.56	40.16	**44.3**(↑)



Confusion matrix of SIB on the Dslr data set

Confusion matrix of SIB on Webcam data set

Figure 6: Confusion matrix of SIB on the Dslr and Webcam datasets.

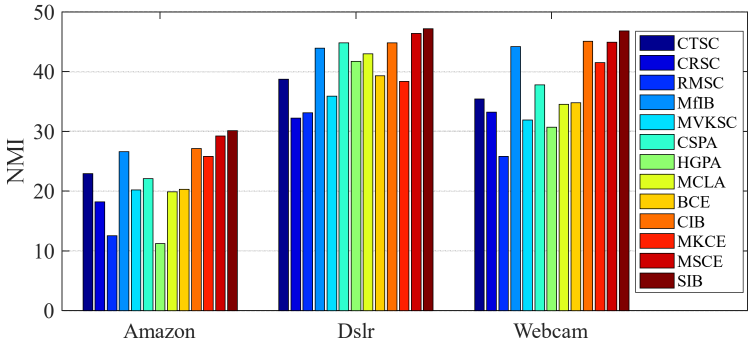


Figure 7: Comparison of the NMI (%) of SIB with those of multi-view and ensemble clustering methods when applied to object category discovery in images.

(31.2%, 25.2%, 30.6% and 22.9%, respectively) against all considered algorithms, mainly because the proposed SIB algorithm can effectively exploit the complementary effect of the multiple feature variables and auxiliary basic clusterings.

5.2.2. *Comparisons with Multi-view and Ensemble Clustering*

This part describes the comparison of the proposed SIB algorithm with the multi-view and ensemble clustering approaches in the context of the task of publication and multilingual corpus analysis. To obtain the basic clusterings of the other ensemble clustering approaches, the original IB was performed 15 times for each feature representation, and the diversity of multiple clusterings was ensured by performing different initializations.

From Table 3, we can observe that the SIB demonstrates better performances than those of the other multi-view and ensemble clustering approaches on the three document datasets. The same observation in terms of NMI can be noted from Fig. 5. Thus, is can be concluded that the proposed SIB algorithm can effectively cope with multiple feature variables and auxiliary basic clusterings. This is mainly because the proposed SIB algorithm can alleviate the overreliance of ensemble clustering methods on existing partitions and mitigate the conflict between heterogeneous features in multi-view clustering.

Table 6: Comparison of the AC (%) of SIB with those of other state-of-the-art image clustering methods.

Datasets	EP	LDMGI	CC	SIB
Amazon	24.8	23.6	30.4	**32.9**(↑)
Dslr	36.2	37.5	47.3	**50.4**(↑)
Webcam	32.5	41.5	47.8	**49.7**(↑)
Average	31.2	34.2	41.8	**44.3**(↑)

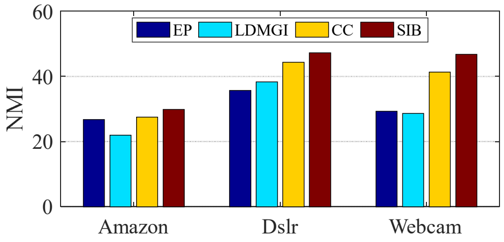


Figure 8: Comparison of the NMI (%) of SIB with those of other state-of-the-art image clustering methods.

5.3. Object Category Discovery in Images

This section describes the application of the SIB to the task of object category discovery in images. In particular, we selected the Dense-SIFT and SURF as the original feature variables; the remaining two features were utilized to construct the auxiliary basic clusterings. Thus, this task involves two original feature variables and two auxiliary basic clusterings.

5.3.1. Experimental Results and Analysis

Table 4 presents the results of the SIB with the IB and other typical clustering methods. From this table, we can observe that (1) the performances of the IB algorithm on the SURF feature are consistently better than those pertaining to the other three features, which demonstrates that the SURF is a discriminative representation of the images. (2) An improper combination of multiple features tends to deteriorate the clustering performance. For instance, although the AC value of the IB when considering the concatenated features improves slightly on the Dslr and Webcam datasets (1.3% and 2.4%, respectively) in comparison with the best value of the IB on a single feature, there is a sharp drop (9.6%) pertaining to the Amazon data. (3) Compared with the typical clustering methods, the average AC values of the SIB method applied to the three image datasets exhibit considerable improvements (18.9%, 15.9%, 17.0% and 16.9%, respectively). This phenomenon demonstrates the effectiveness of the SIB on the task of object category discovery in images. To further demonstrate the superiority of the SIB, we visualized the confusion matrices of the SIB for the Dslr and Webcam, as shown in Fig. 6. This figure shows that the learned object categories are relatively pure, and each of them can be highly correlated with the true cluster label.

Table 7: Comparison of the AC (%) of SIB with those of the original IB and the other four typical clustering methods when applied to unsupervised human action categorization in videos.

Datasets	IB				IB	*k*-means	pLSA	LDA	NCuts	SIB
	HOG	HOF	STIP	3DSIFT	Con					
UCF Sports	38.7	53.8	50.4	37.1	53.7	40.3	46.3	51.7	47.1	**59.6**(↑)
UCF 50	33.1	34.0	31.2	33.3	33.9	29.0	30.3	29.6	31.7	**40.2**(↑)
HMDB	19.0	22.3	21.8	28.1	26.3	21.4	22.4	23.3	23.2	**29.8**(↑)
Average	30.3	36.7	34.5	32.8	38.0	30.2	33.0	34.9	34.0	**43.2**(↑)

Table 8: Comparison of the AC (%) of SIB with those of the multi-view and ensemble clustering methods when applied to unsupervised human action categorization.

Datasets	Multi-view clustering					Ensemble clustering							SIB
	CTSC	CRSC	RMSC	MfIB	MVKSC	CSPA	HGPA	MCLA	BCE	CIB	MKCE	MSCE	
UCF Sports	53.6	46.3	49.2	54.0	46.2	49.2	56.8	53.0	51.5	56.9	40.65	51.96	**59.6**(↑)
UCF 50	35.0	31.5	32.6	36.2	35.8	32.0	26.2	32.9	30.1	36.8	31.82	36.21	**40.2**(↑)
HMDB	24.5	25.8	22.2	26.5	25.3	27.0	20.6	22.3	20.8	27.2	15.81	28.37	**29.8**(↑)
Average	37.7	34.5	34.7	38.9	35.8	36.1	34.5	36.0	34.1	40.3	29.43	38.85	**43.2**(↑)

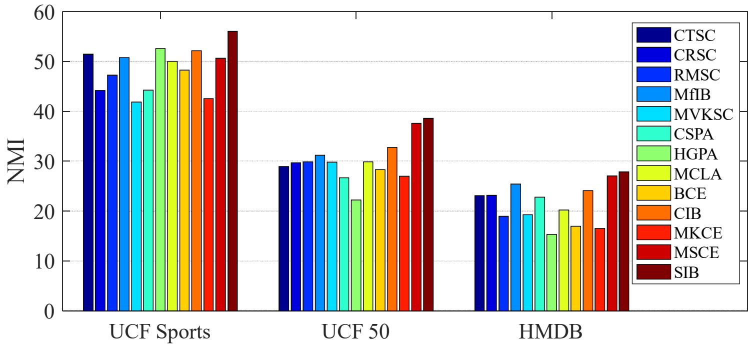


Figure 9: Comparison of the NMI (%) of SIB with those of the multi-view and ensemble clustering methods when applied to unsupervised human action categorization in videos.

Table 5 and Fig. 7 shows the comparison results of the SIB with other multi-view and ensemble clustering methods when applied to the task of object category discovery in images. From this table, we can observe that (1) the multi-view and ensemble clustering methods demonstrate improved clustering performances compared with those of the single-view clustering algorithms (in Table 4). (2) The AC values of the SIB algorithm are consistently considerably better than those of the other multi-view and ensemble clustering approaches. We can conclude that the proposed SIB algorithm can effectively deal with multiple feature variables and auxiliary basic clusterings when applied to the task of object category discovery in images.

5.3.2. Comparison with State-of-the-art Image Clustering Methods

To enable a comparison of the SIB algorithm with other state-of-the-art image clustering methods, we adopted local discriminant models and global integration (LDMGI) [53], clustering-by-composition (CC) [54] and ensemble projection (EP) [55] as baselines. LDMGI learns a new Laplacian matrix by using both the manifold structure and local discriminant information, which makes it more robust for data clustering. CC is based on the composition of an image from large non-trivial pieces of other images in which similar images can be easily composed from each other. EP learns a new feature representation by capturing the information of each image and the relatedness across images. As shown in Table 6 and Fig. 8 the SIB method demonstrates improvements on all the data sets compared with the other three state-of-the-art image clustering methods. The average AC values of the SIB on the three image data sets demonstrate an improvement of 13.1%, 10.1%, and 2.5%, respectively.

5.4. Unsupervised Human Action Categorization

We applied the SIB to the task of human action categorization in an unsupervised setting. Specifically, we selected Dense-SIFT and SURF as the original feature variables, and the remaining two features were utilized to construct the auxiliary basic clusterings. Thus, this task involved two original feature variables and two auxiliary basic clusterings.

Table 9: Comparison of the AC (%) of SIB with those of other state-of-the-art action clustering methods.

Datasets	DAKM	MvIB	CIB	SIB
UCF Sports	53.9	55.3	56.9	**59.6**(↑)
UCF 50	34.5	36.1	36.8	**40.2**(↑)
HMDB	26.3	27.5	27.2	**29.8**(↑)
Average	38.2	39.6	40.3	**43.2**(↑)

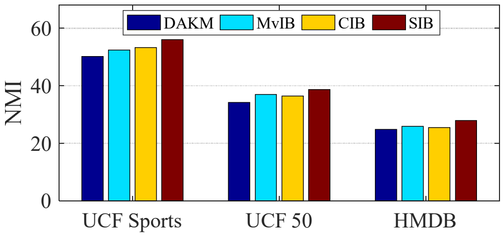


Figure 10: Comparison of the NMI (%) of SIB with those of other state-of-the-art action clustering methods.

5.4.1. Experimental Results and Analysis

Table 7 presents the experimental results of the SIB compared with the original IB and other typical clustering methods. From this table, the following observations can be made: (1) The original IB algorithm performs differently on HOG, HOF, STIP and 3DSIFT features. For instance, the IB algorithm gets the best result when using the 3DSIFT feature (28.1%) on HMDB dataset, while the best AC is obtained when considering the HOF feature (53.8%) on the UCF Sports dataset. This phenomenon also verifies that a single feature is not sufficiently discriminative for the action categorization in different video datasets. (2) When concatenating multiple features together, the original IB algorithm does not demonstrate improved performances consistently on the video datasets. For instance, the performance declines on the HMDB dataset (1.8%) compared with the best result of the IB on a single feature. (3) The SIB method demonstrates considerable improvements (13.2%, 10.2%, 8.3% and 9.2%, respectively) in terms of the average results compared with those of k-means, pLSA, LDA and NCuts algorithms. It is clear that the performances of the SIB are consistently better than those of the original IB and other typical clustering methods.

Table 8 and Fig. 9 show the AC and NMI results of the SIB algorithm compared with other multi-view and ensemble clustering approaches. The following observations can be made considering these results: (1) the multi-view and ensemble methods demonstrate certain clustering performance improvements over the single-view clustering algorithms (see Table 7); and (2) the AC values of the SIB are higher than those of the other multi-view and ensemble clustering approaches. The NMI values shown in Fig. 9 corroborate this observation, which again reflects that the proposed SIB algorithm can effectively deal with multiple feature variables and auxiliary basic clusterings when applied to unsupervised human action categorization in videos.

5.4.2. Comparison with Action Clustering Methods

In this section, three state-of-the-art action clustering methods, namely, the dual assignment k-means (DAKM) [36], multivariate video information bottleneck (MvIB) [56] and the preliminary version of this work named consensus information bottleneck (CIB) [32], are employed to verify the effectiveness of the SIB method when applied to the task of unsupervised human action categorization.

5.5. Parameter Analysis

Table 9 lists the AC value of the SIB compared with those of the action clustering methods. The table indicates that the proposed SIB algorithm performs better than the other action clustering algorithms on the three video datasets. Specifically, the average results of SIB on UCF Sports, UCF 50 and HMDB indicate improvements (5.0%, 3.6% and 2.9%, respectively) compared with DAKM, MvIB and CIB. Fig. 10 shows the comparison of the NMI results of the

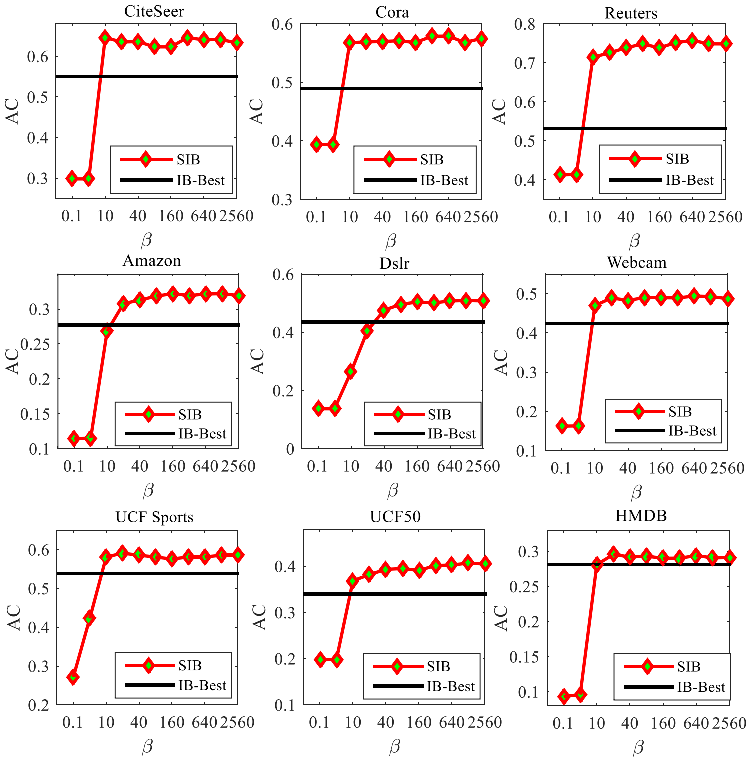


Figure 11: Performances of SIB with various β

SIB with those of other action clustering methods. The same observation in terms of the AC value can be seen from the figure. Thus, it can be concluded that the proposed SIB method can effectively deal with the task of unsupervised human action categorization in videos as well.

The proposed SIB generally formulates the problem of joint multi-view and ensemble clustering as a function of the mutual information maximization. In this objective function, the information filtered from the original features and auxiliary basic clusterings is maximally preserved through a “bottleneck” with respect to the final clustering partition, while the source data are compressed into its clustering partition as much as possible. As we can see from function (6), SIB utilizes β to strike the balance between the data compression and relevant information preservation. This section describes the analysis of the impact of trade-off parameter β on the performance of SIB on all datasets considered in this study. In particular, we vary the values of β from the value set {0.1, 1, 10, 20, 40, 80, 160, 320, 640, 1280, 2560}. From Fig. 11, we can make the following observations: First, when $\beta \rightarrow 0$, SIB performs poorly since it considers only the compression of the source data X to its clustering partition T. When the value of β is increased, the performance of the SIB is considerably better better because it strikes a balance between the data compression and information preservation. This fact also explains why many IB applications set β as ∞ [27, 31]. In this study, we set the β as 80 on all the data sets.

5.6. *Convergence Analysis*

As mentioned in the section on theoretical analysis, the SIB algorithm can converge in a few iterations. This section describes the empirical testing of the convergence of the SIB. As we can see from Fig. 12, every repetition increases the values of objective function 6 and 30 iterations are sufficient for convergence.

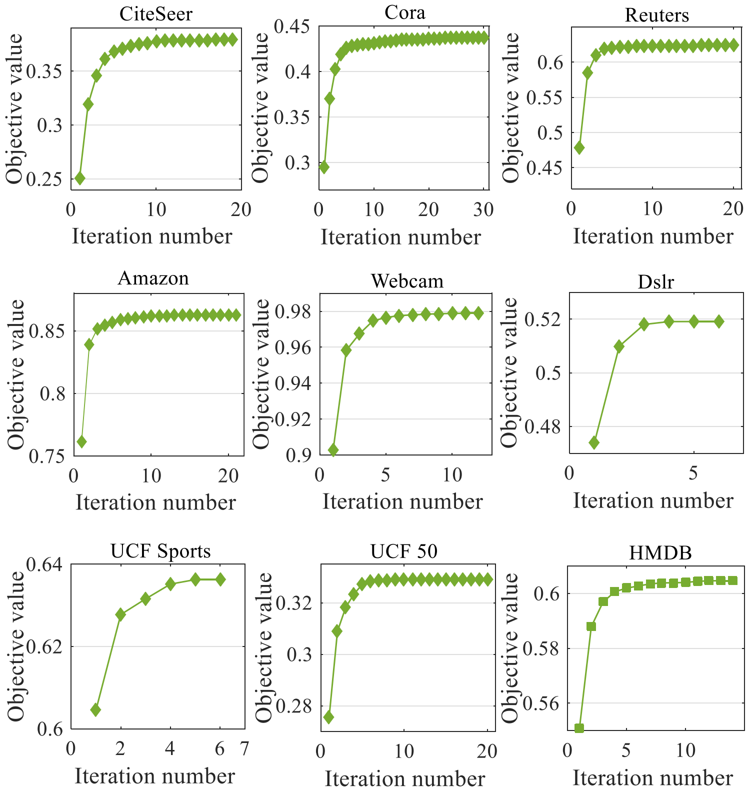


Figure 12: The iterations of SIB

6. Conclusion

In this work, we performed simultaneous multi-view and ensemble clustering jointly by extending the information bottleneck theory into a novel synergetic information bottleneck (SIB) method. The SIB determines the final data partition by considering the original features and auxiliary base clusterings simultaneously, in which the original features characterize data information from different views, while the base clusterings reveal the data information from heterogeneous features. Specifically, SIB generally formulates the problem of joint multi-view and ensemble clustering as a function of mutual information maximization. In this function, the information between original features and basic auxiliary clusterings is preserved simultaneously in terms of the final clustering partition. In addition, to solve the optimization of SIB objective function, a sequential draw-and-merge optimization solution is presented to update the data partition. The experiments pertaining to the task of publication and multilingual corpus analysis, object category discovery in images and unsupervised human action categorization have confirmed the effectiveness of the proposed SIB algorithm.

7. Acknowledgements

This work was supported by the National Natural Science Foundation of China under grant No. 61772475 and 61906172, the National Key Research and Development Program of China under grant No. 2018YFB1201403, and the EPSRC through project 4D Facial Sensing and Modelling under grant No. EP/N025849/1.

References

[1] C. Xu, D. Tao, C. Xu, Large-margin multi-view information bottleneck, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 36 (8) (2014) 1559–1572.

[2] K. Chaudhuri, S. M. Kakade, K. Livescu, K. Sridharan, Multi-view clustering via canonical correlation analysis, in: International Conference on Machine Learning (ICML), 2009, pp. 129–136.

[3] A. Kumar, H. Daumé, A co-training approach for multi-view spectral clustering, in: International Conference on Machine Learning (ICML), 2011, pp. 393–400.

[4] A. Kumar, P. Rai, H. Daume, Co-regularized multi-view spectral clustering, in: Advances in Neural Information Processing Systems (NIPS), 2011, pp. 1413–1421.

[5] H. Wang, C. Weng, J. Yuan, Multi-feature spectral clustering with minimax optimization, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4106–4113.

[6] X. Cao, C. Zhang, H. Fu, S. Liu, H. Zhang, Diversity-induced multi-view subspace clustering, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 586–594.

[7] C. Wang, J. Lai, S. Y. Philip, Multi-view clustering based on belief propagation, IEEE Transactions on Knowledge and Data Engineering (TKDE) 28 (4) (2016) 1007–1021.

[8] L. Houthuys, R. Langone, J. A. K. Suykens, Multi-view kernel spectral clustering, Information Fusion 44 (2018) 46–56.

[9] X. Cai, F. Nie, W. Cai, H. Huang, Heterogeneous image features integration via multi-modal semi-supervised learning model, in: IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1737–1744.

[10] Z. Zhang, L. Liu, F. Shen, H. T. Shen, L. Shao, Binary multi-view clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) Early Access (99) (2018) 1–9.

[11] A. Strehl, J. Ghosh, Cluster ensembles–a knowledge reuse framework for combining multiple partitions, Journal of Machine Learning Research (JMLR) 3 (2002) 583–617.

[12] X. Z. Fern, C. E. Brodley, Solving cluster ensemble problems by bipartite graph partitioning, in: International Conference on Machine Learning (ICML), 2004, pp. 36–43.

[13] A. L. Fred, A. K. Jain, Combining multiple clusterings using evidence accumulation, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 27 (6) (2005) 835–850.

[14] T. Li, C. Ding, M. Jordan, et al., Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization, in: IEEE International Conference on Data Mining (ICDM), 2007, pp. 577–582.

[15] N. Iam-On, T. Boongoen, S. Garrett, C. Price, A link-based approach to the cluster ensemble problem, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 33 (12) (2011) 2396–2409.

[16] H. Wang, H. Shan, A. Banerjee, Bayesian cluster ensembles, Statistical Analysis and Data Mining (SADM) 4 (1) (2011) 54–70.

[17] A. Loureno, S. R. Bul, N. Rebagliati, A. L. N. Fred, M. A. T. Figueiredo, M. Pelillo, Probabilistic consensus clustering using evidence accumulation, Machine Learning (ML) 98 (2) (2015) 331–357.

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

[19] P. Zhou, L. Du, H. Wang, L. Shi, Y.-D. Shen, Learning a robust consensus matrix for clustering ensemble via kullback-leibler divergence minimization, in: International Joint Conference on Artificial Intelligence (IJCAI), 2015, pp. 4112–4118.

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

[21] D. Huang, C. D. Wang, J. H. Lai, Locally weighted ensemble clustering, IEEE Transactions on Cybernetics (TCYB) 48 (5) (2018) 1460–1473.

[22] J. Zhao, X. Xie, X. Xu, S. Sun, Multi-view learning overview: Recent progress and new challenges, Information Fusion 38 (2017) 43–54.

[23] E. Bruno, S. Marchand-Maillet, Multiview clustering: A late fusion approach using latent models, in: International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2009, pp. 736–737.

[24] Q. Zhang, S. Sun, Multiple-view multiple-learner active learning, Pattern Recognition (PR) 43 (9) (2010) 3113–3119.

[25] X. Xie, S. Sun, Multi-view clustering ensembles, in: International Conference on Machine Learning and Cybernetics (ICMLC), 2013, pp. 51–56.

[26] Z. Tao, H. Liu, S. Li, Z. Ding, Y. Fu, From ensemble clustering to multi-view clustering, in: International Joint Conference on Artificial Intelligence (IJCAI), 2017, pp. 2843–2849.

[27] N. Tishby, F. C. Pereira, W. Bialek, The information bottleneck method, in: Annual Allerton Conference on Communnication, Control and Computing, 1999, pp. 368–377.

[28] N. Slonim, N. Friedman, N. Tishby, Multivariate information bottleneck, Neural Computation 18 (8) (2006) 1739–1789.

[29] T. M. Cover, J. A. Thomas, Elements of Information Theory, Birkhäuser Basel, 1991.

[30] Z. Lou, Y. Ye, X. Yan, The multi-feature information bottleneck with application to unsupervised image categorization, in: International Joint Conference on Artificial Intelligence (IJCAI), 2013, pp. 1508–1515.

[31] X. Yan, S. Hu, Y. Ye, Multi-task clustering of human actions by sharing information, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4049–4057.

[32] X. Yan, Y. Ye, X. Qiu, Unsupervised human action categorization with consensus information bottleneck method, in: International Joint Conference on Artificial Intelligence (IJCAI), 2016, pp. 2245–2251.

[33] G. Chechik, N. Tishby, Extracting relevant structures with side information, in: Advances in Neural Information Processing Systems (NIPS), 2002, pp. 857–864.

[34] D. Gondek, T. Hofmann, Conditional information bottleneck clustering, in: IEEE International Conference on Data Mining (ICDM), 2003, pp. 36–42.

[35] M. Rey, V. Roth, T. J. Fuchs, Sparse meta-gaussian information bottleneck, in: International Conference on Machine Learning (ICML), 2014, pp. 910–918.

[36] S. Jones, L. Shao, Unsupervised spectral dual assignment clustering of human actions in context, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 604–611.

[37] Y. Ye, R. Liu, Z. Lou, Incorporating side information into multivariate information bottleneck for generating alternative clusterings, Pattern Recognition Letters (PRL) 51 (2015) 70–78.

[38] N. Slonim, The information bottleneck: Theory and applications, Ph.D. thesis (2002).

[39] H. Bay, A. Ess, T. Tuytelaars, L. J. V. Gool, Speeded-up robust features (SURF), Computer Vision and Image Understanding (CVIU) 110 (3) (2008) 346–359.

[40] L. Wolf, T. Hassner, Y. Taigman, Descriptor based methods in the wild, in: Workshop on Faces in Real Life Images Detection Alignment and Recognition, 2008.

[41] F. S. Khan, J. van de Weijer, M. Vanrell, Top-down color attention for object recognition, in: IEEE International Conference on Computer Vision (ICCV), 2009, pp. 979–986.

[42] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2006, pp. 2169–2178.

[43] M. D. Rodriguez, J. Ahmed, M. Shah, Action mach a spatio-temporal maximum average correlation height filter for action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.

[44] K. K. Reddy, M. Shah, Recognizing 50 human action categories of web videos, Machine Vison and Applications (MVA) 24 (5) (2013) 971–981.

[45] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, Hmdb: a large video database for human motion recognition, in: IEEE International Conference on Computer Vision (ICCV), 2011, pp. 2556–2563.

[46] I. Laptev, On space-time interest points, International Journal of Computer Vision (IJCV) 64 (2) (2005) 107–123.

[47] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 886–893.

[48] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: ACM International Conference on Multimedia (ACM'MM), 2007, pp. 357–360.

[49] T. Hofmann, Probabilistic latent semantic analysis, in: Uncertainty in Artificial Intelligence (UAI), 1999, pp. 391–407.

[50] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research (JMLR) 3 (2003) 993–1022.

[51] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on Pattern Analysis Machine Intelligence (TPAMI) 22 (8) (2000) 888–905.

[52] R. Xia, Y. Pan, L. Du, J. Yin, Robust multi-view spectral clustering via low-rank and sparse decomposition, in: American Association for Artificial Intelligence (AAAI), 2014, pp. 2149–2155.

[53] Y. Yang, D. Xu, F. Nie, S. Yan, Y. Zhuang, Image clustering using local discriminant models and global integration, IEEE Transactions on Image Processing (TIP) 19 (10) (2010) 2761–2773.

[54] A. Faktor, M. Irani, Clustering by composition-unsupervised discovery of image categories, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 36 (6) (2014) 1092–1106.

[55] D. Dai, L. J. V. Gool, Unsupervised high-level feature learning by ensemble projection for semi-supervised image classification and image clustering, CoRR abs/1602.00955.

[56] X. Yan, Y. Ye, Z. Lou, Unsupervised video categorization based on multivariate information bottleneck method, Knowledge-Based Systems (KBS) 84 (2015) 34–45.

[57] D. Cai, X. Wang, X. He, Probabilistic dyadic data analysis with local and global consistency, in: International Conference on Machine Learning (ICML), 2009, pp. 105–112.