Unsupervised Cross-Language Classification with Stratified Sampling-Based Cluster Ensemble

Wenli Gui, Liping Jing, Liu Yang, and Jian Yu

***Abstract*—Many real world data sets are comprised of multiple representations or views, learning from multi-view data is important in many applications. In the unsupervised cross-language classification problems, the documents in different languages always share the same set of categories. To solve the cross-language clustering problem, we propose a novel Stratified Sampling-based Cluster Ensemble method, which has two main contributions. It can effectively generate several data components from the cross-language documents set via stratified sampling technique, so that the correlation between multiple views can be significantly considered. On the other hand, it makes use of the linked based consensus function to combine the component clustering results, so that the relationship between components can be effectively utilized.**
A series of experiments on real cross-language documents set have been conducted. The experimental results have shown that the proposed method outperforms the state-of-the-art multi-view clustering methods.

***Index Terms*—Unsupervised cross-language classification, multi-view clustering, clustering ensemble, stratified sampling.**

I. Introduction

With the rapid growth of global processing, cross-language data have been used in most aspects of human society. It is very common that documents in different languages share the same set of categories. In the real world, many data sets are naturally comprised of multiple views. For instance, web pages can be represented by both the page-text and the hyperlinks pointing to them, which form two independent views [1]. In natural language processing tasks, the same story can be described in articles from different news sources, and one document can have multiple representations in multiple different languages. In these applications, although to some extent, each individual view can characterize the data object, the multiple views can contain much more complementary information and knowledge to each other to alleviate the difficulty of a giving learning task. Multi-view learning becomes a common theme that exploiting multiple redundant views to effectively learn from the data and improve the performance of the target learning task [2].

Much work has been done for multi-view learning, however, most of the methods were focused on classification problems [3]-[5]. As we all know, the true labels of the data sets are very hard to obtain. It is expensive to obtain the accurate labels from domain experts. Thus, it is urgent to automatically identify the label information for the large volume of data.

Recently, it has gained increasing attention to exploiting multiple views to improve unsupervised learning from machine learning research community [2]. A number of multi-view clustering methods have been developed in the literature such as the two-view spectral clustering over bipartite graphs method [6], the canonical correlation analysis (CCA) method [7], the generalized multi-view normalized cut method [8], and the Multi-NMF method [9]. All these methods suggest that learning low-dimensional representations consistent across multiple views can improve the clustering performance. Nevertheless, the first two unsupervised multi-view learning methods are limited because they focused only on two-view learning problems. Although the third one, Multi-NMF [9], has ability to deal with multiple views problem, there are some disadvantages in this algorithm, for example, too many parameters to be predefined.

Previous methods on unsupervised multi-view learning are limited to two-view learning, or they ignore the relationship between all the views and need parameter knowledge. In order to solve the above problems, we propose a novel stratified sampling-based cluster ensemble method for unsupervised cross-language classification (SSCE-CLC), which has two main contributions. It can effectively generate several data components from the cross language documents set via stratified sampling technique, so that the correlation between different languages (i.e., multiple views) can be significantly considered. On the other hand, it makes use of the linked based consensus function to combine the component clustering results, so that the relationship between components can be effectively utilized.

The paper is organized as follows: Section II states the background of cross-language clustering. Section III describes our main method for unsupervised cross-language classification. Section IV presents the experimental results of our proposed method and the discussion of the results. Section V gives a brief conclusion and future work.

II. Background

To make this paper self-contained, we first provide some background knowledge and introduce some notations which

Manuscript received October 15, 2014; revised December 16, 2014. This work was supported in part by the NSFC （61375062 and 61370129), the Ph.D Programs Foundation of Ministry of Education of China (20120009110006), the Opening Project of State Key Laboratory of Digital Publishing Technology, the Fundamental Research Funds for the Central Universities (2014JBM029 and 2014JBZ005), and the Program for Changjiang Scholars and Innovative Research Team (IRT 201206).

Wenli Gui, Liping Jing, and Jian Yu are with Beijing Key Lab of Traffic Data Analysis and Mining, the School of Computer Science and Information Technology, Beijing Jiaotong University, Beijing 100044 China (e-mail: 13125158@bjtu.edu.cn, lpjing@bjtu.edu.cn, jyu@bjtu.edu.cn).

Liu Yang is with Beijing Key Lab of Traffic Data Analysis and Mining, the School of Computer Science and Information Technology, Beijing Jiaotong University, Beijing 100044 China, and College of Mathematics and Computer Science, Hebei University, Baoding, Heibei, China (e-mail: 11112091@bjtu.edu.cn).

be used throughout the rest of the paper. For unsupervised cross language classification problem, we take one language as a view. Given a cross language data set, i.e., English ($X^{(1)}$), French ($X^{(2)}$), ..., Spanish ($X^{(M)}$) and so on, each view has different dimensional feature space, i.e. $F^{(1)}, F^{(2)}, \dots, F^{(M)}$. When we take clustering algorithm to single view, the performance may be not very good. As we know, languages as the vectors for communication, they are related with each other. Therefore, we want to take the relationship between all the languages into account for improving the performance of multi-view clustering.

A. Non-negative Matrix Factorization

In the primary notation part, we briefly introduce the Non-Negative Matrix Factorization (NMF) [10]. Let $X = [X_{.,1}, \dots, X_{.,N}] \in R_+^{M*N}$ denote the nonnegative data matrix where each column represents a data point and each row represents one attribute. NMF want to find two nonnegative matrix factors $U = [U_{i,k}] \in R_+^{M*K}$ and $V = [V_{j,k}] \in R_+^{N*K}$ whose product provides a good approximation to X:

$$X \approx UV^T \quad (1)$$

Here *K* denotes the desired reduced dimension. In this case, the original data X can be deduced into a low-dimensional representation V with the aid of basis matrix U. In the view of clustering, U can be taken as the cluster centers, while V indicates the clustering coefficient between the data points and the corresponding clusters. NMF has been proved to be an effective and efficient clustering algorithm in real applications, esp., for text clustering [9], [11].

B. Multi-view Clustering Algorithm

Much work has been done in multi-view clustering, but some of them limit to two views task. We concentrate our attention to multiple views (more than two views) clustering learning. For unsupervised multi-view learning, there already exist many methods such as Collective NMF was proposed in [11] and Multi-view NMF in [9].

Collective NMF used the shared coefficient matrix but different basis matrices across views as shown below:

$$\sum_{i=1}^{M} \lambda_i \left\| X^{(i)} - U^{(i)}(V^{(*)})^T \right\|_F^2 \quad (2)$$

This method assumed that different views should exactly share a same clustering results *V**. However, in real application, there may be some noise or disturbed information in data, such assumption may result in bad result. In order to solve this problem, Liu *et al.* [9] proposed a multi-view NMF model as follows.

$$\sum_{v=1}^{M} \left\| X^{(v)} - U^{(v)}\left(V^{(v)}\right)^T \right\|_F^2 + \sum_{v=1}^{M} \lambda_v \left\| V^{(v)} - V^* \right\|_F^2 \quad (3)$$
$$\text{s.t.} \forall 1 \le k \le K, \left\| U_{.,k}^{(v)} \right\|_1 \text{and } U^{(v)}, V^{(v)}, V^* \ge 0$$

There are M parameters λ_v to control the effect of all views. In real applications, it is hard to predefine a proper value for so many parameters.

C. Stratified Sampling Method

Stratified Sampling is well known in sampling techniques for it can capture the population characteristics adequately [12]. The Stratified Sampling method is conducted in two steps, firstly it dividing the whole population into subpopulations, and then we applying random sampling to each subpopulations to select the representative samples. Stratified Sampling method is widely used in many areas, such as web mining [13], traffic data analysis [14] and random forest for classification [15].

D. Clustering Ensemble

In order to make up for the problem that one single clustering algorithm can not perform very well on a given data set, clustering ensemble is attractive to cluster high dimensional data such as text data, microarray data and image data [16]. It integrates multiple clustering results generated from samples of a given data set into a single clustering with a result which is usually much better than the results of individual clustering on the data set. Given a data set, the process of clustering ensemble is conducted in two steps, generating a set of individual clustering results from the data set and integrating the component results into a clustering ensemble.

For cross-language, the documents set can be taken as a multi-view data set, each language refers to one view. It is intuitive that taking the data in one view as a component data. This motivates us to find a method to combine the component clustering results generated from the multiple views, and then apply the clustering ensemble to analyze the cross-language data.

However, such simple clustering ensemble strategy ignores the relationship between multiple views. Thus, we proposed a Stratified Sampling-based Clustering Ensemble method to effectively generate clustering components from the cross-language documents, and then combine the component results to obtain the final clustering result. More details will be given in next section.

III. Methods

A. Proposed Learning Framework

Given a cross language data set for unsupervised multi-view learning, let $X^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \dots x_N^{(i)}\}$ (*i*=1, 2, … , *M*) be a set of *N* data points of M views. As we know, the data sets of *M* views have their own feature spaces. Each view has its own feature space $F^{(i)} = \{f_1, f_2, \dots, f_{d^{(i)}}\}$ (*i*=1, 2, … , *M*) which $d^{(i)}$ is the feature dimension of the data set $X^{(i)}$. $X^{(i)}$ is denonted in document-feature matrix, in the document-feature matrix, each row represents a document, each column represents a feature, and the cell contains the weighted value of a feature for a document. For clustering, we assume that a data point in different views would be assigned to the same cluster with high probability. In our latter experiment, we apply Document Frequency (DF) to the real world multi-view data sets.

Previous cross language clustering methods need predefined parameters [9] or ignore the relationship between multiple views [16]. As we know, the relationship between different languages is important to help understanding the semantics of cross-language documents. Therefore, we propose a novel framework by taking advantage of the relationship of all views as shown in Fig. 1. More details are given as follows.

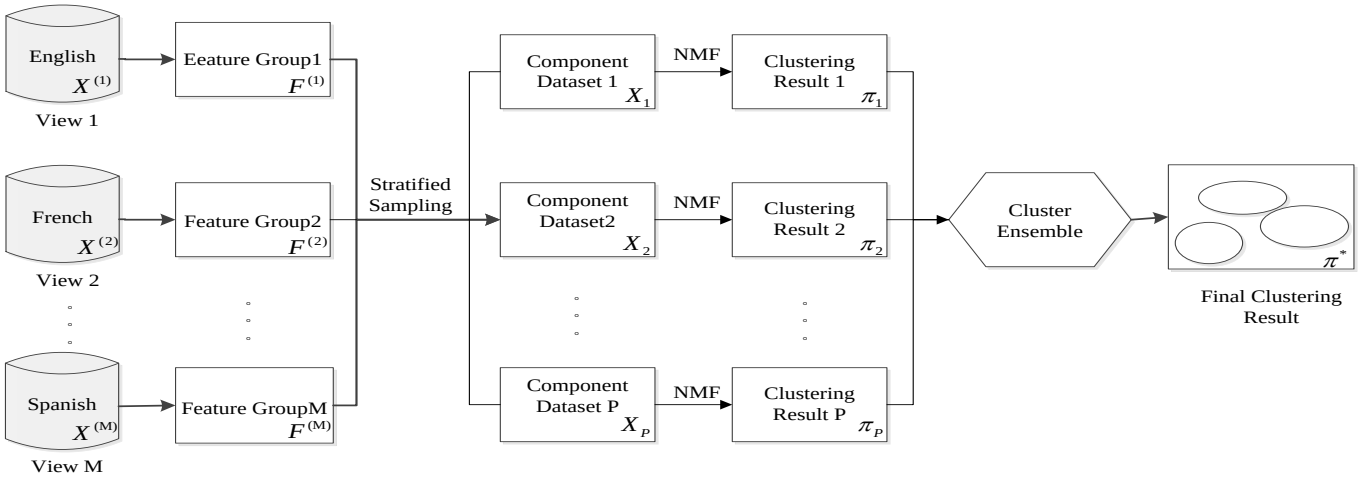

Fig. 1. The framework of Stratified Sampling-based Cluster Ensemble for unsupervised cross-language classification.

Given a multi-view data set $X^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \dots, x_N^{(i)}\}$ (i=1, 2, … , M) with a set of features $F^{(i)} = \{f_1, f_2, \dots, f_{d^{(i)}}\}$ (i=1, 2,…, M), we want to generate P component data sets $(X_1, X_2, \dots, X_P)$ with $d^{(i)}$ dimensions by sampling X on F. How to sample the features from the data sets of multiple views is a difficult but significant problem for us. We calculate the total number of features from all views, note to T. According to the latter experiments, we choose 15% as the best ratio to select the number of samples, note to $Q = 15\% * T$. We calculate the variance of all views, and then we get the percentage to select features from every view. The percentage of every view to sampling is

$$P\left(F^{(i)}\right) = \frac{var\left(F^{(i)}\right)}{\sum_{i=1}^{M} var\left(F^{(i)}\right)} \quad (i = 1, \dots, M) \qquad (4)$$

We use the variance to decide the ratio to sampling feature from every view. Due to the variance reflect the typicality of each view. The following is the formula to calculate the component

$$\mathrm{D}^{(p)} = \sum_{i=1}^{M} \mathrm{Q} * P\left(F^{(i)}\right) \qquad (5)$$
$$(p = 1, \dots, P;\ \mathrm{i} = 1, \dots, M)$$

where $\mathrm{D}^{(p)}$ is the dimension size of the data component p which be constructed by using stratified samplings. We generate P components data set $X_{(i)} = \{X_1, X_2, \dots, X_p\}$ (i=1, 2, … , P). After generate P components data set, we utilize NMF clustering algorithm to generate P clustering results, i.e. clustering result 1, …, clustering result P, we note to $\amalg = \{\pi_1, \pi_2 \dots \dots \pi_P\}$. Then we take cluster ensemble method for the P components. $\amalg = \{\pi_1, \pi_2 \dots \dots \pi_P\}$ is a set of P base clustering results, which is referred to as a cluster ensemble. Each base clustering results (ensemble member) returns a set of clusters $\pi_i = \{C_1^i, C_2^i \dots \dots C_{k_i}^i\}$, such that $\bigcup_{j=1}^{k_i} C_j^i = X^{(i)}$, where k_i is the number of clusters in the i-th clustering. For each $x \in X^{(i)}$, $C(x)$ denotes the cluster label to which the data point x belongs. In the i-th clustering, C(x) = j if $x \in C_j^i$. We want to find a new clustering result π^* of the data set $X^{(i)}$ that summarizes the information from the cluster ensemble $\amalg$.

Accordingly, the figure clearly shows us there are two steps of our Stratified Sampling-based Cluster Ensemble method: 1) generate the base clustering components from multi-view data, 2) produce the final clustering result for multi-view data using consensus function. In our multi-view clustering problem, we use the same clustering algorithm NMF for generating base clustering results. And we utilize the link-based consensus function to generate the final clustering result for multi-view data.

B. Component Generation from Multi-View Data

Taking into account the relationship between the multiple views, we propose to take a cluster ensemble method to deal with the multi-view clustering problem. But how to generate the components to ensemble and how to consider the relationship between all the views is still a important task for us. We apply the stratified sample method to generate the components. For data sets $X^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \dots, x_N^{(i)}\}$ (i=1, 2, …, M) , every view has its own feature space, but they are related with each other. If we want to utilize the relationship between the features of every view, we should look for a method to fuse the features of every view. Therefore, we utilize sampling method to select features. Considering the effectiveness and accuracy of the sampling method, we choose stratified sampling method as our sampling method.

Firstly, we note the features of each view to a set $F^{(i)} = \{f_1, f_2, \dots, f_{d^{(i)}}\}$ ($i = 1, \dots, M$), $F^l \cap F^t = \emptyset$ $(1 \le l, t \le M, l \ne t)$. The number of features in $F^{(i)}$ is denoted as $d^{(i)}$, i.e., $\left|F^{(i)}\right| = d^{(i)}$. To generate the component data set $X_{(i)}$ (i= 1, 2, …, P), we randomly sample feature subset with $Q * P(F^{(i)})$ features from each feature set $F^{(i)} = \{f_1, f_2, \dots, f_{d^{(i)}}\}$ (i=1, 2,…, M).

To make the features we choose from every view can represent the view, we make the features in each subset $F^{(i)}$ positively correlated and similar to each other as much as possible. Let $\{F^{(1)}, F^{(2)}, \dots, F^{(M)}\}$ be a set of M feature groups or strata and $d^{(i)}$ (i=1, 2,…, M) is the number of features in the feature group $F^{(i)}$. Let T be the number of features in $X^{(i)}$ and $T = d^{(1)} + d^{(2)} + \cdots + d^{(M)}$. Assume that objects in $X^{(i)}$ are independent from each other. We apply the variance to decide the scale of feature to sample. The scale of the sample from each view is $P(F^{(i)})$ which has been given in formula (4).

We calculate the total number of features from all views, note to *T*. According to the latter experiments for sampling rates, we choose 15% as the percentage to select the number of features, note to $Q = 15\% * T$. To make the features that we select more typically, stratified sampling method is used to our problem. Note that there may be common or overlapping feature among component data sets because any feature has the same chance of selection in every random sampling.

We use the variance to decide the number to sampling from each view. The following formula computes the feature number to sampling from each view.

$$S^{(i)} = Q * P\left(F^{(i)}\right) \ (i = 1, \dots, M) \tag{6}$$

Then we combine the features sampled from each view to an ensemble matrix. We randomly sampling P times, and generate *P* components data set $X_{(i)} = \{X_1, X_2, \dots, X_p\}$ (*i*=1, 2, ..., *P*). After that, we utilize NMF to the component data sets to generate clustering results $\Pi = \{\pi_1, \pi_2 \dots \dots \pi_P\}$(*i*=1, 2, ..., *P*). Therefore, the component data set by stratified sampling is more representative to the whole data set than the component data set by random sampling. Having P clustering results, we want a basic and straightforward method to integrate multiple clusterings into a single clustering. In this representation, objects occurring in the same set of clusters tend to be clustered together in the clustering ensemble.

C. Component Confused for Multi-View Learning

Given a multi-view data set $X^{(i)}$, ensemble clustering of $X^{(i)}$ is a process to integrate multiple clustering results and produced by one or more clustering algorithms from component data sets into a single final clustering result. Ensemble clustering consists of two major steps, generation of component clustering results and integration of component results into an individual clustering result. The framework is illustrate in the Fig. 1. For our multi-view clustering problem, the first step is shown in the previous part. A cross language data sets $X^{(i)}$ is sampled to *P* component data sets $X_{(i)} = \{X_1, X_2, \dots, X_p\}$ (*i*=1, 2, …, *P*) and each component data set is clustered independently to create *P* component results $\{\pi_1, \pi_2 \dots \dots \pi_P\}$. Here, we utilize the NMF clustering as our base clustering algorithm. The second step takes the *P* component results as input and integrates them with a consensus function into a final clustering result.

Having obtained the component clustering results, we need to search a good consensus function to ensemble the clustering results. As we know, clusters in two arbitrary different clustering results are directly related and the relationship between two clusters can be computed with a distance or the similarity measure. Using the relationship information in the integration step is important to improve the accuracy of the clustering ensemble. Therefore, we utilize consensus functions to develop the use of the relationship information to ensemble the P clustering results.

In our Stratified Sampling-based Cluster Ensemble method, we take Link-based consensus function (LB) as the consensus function which was recently proposed in [16]. In the link-based approach, a weighted and undirected hyper graph *G* = (*V*, *W*) has been constructed, where *V* is the set of vertices each representing a cluster produced by X_p, and *W* is a set of weighted edges between clusters. The similarity of two vertices is estimated by counting the number of Connected-Triples they are part of and regarding each triple as the minimum weight of the two involving edges. Specifically, the similarity between two clusters C_i, C_j is defined as

$$\mathrm{Sim}\left(C_i, C_j\right) = \alpha \frac{CT_{ij}}{CT_{max}} \tag{7}$$

$$CT_{ij} = \sum_{k=1}^{t} \min\left(\frac{D_i \cap D_k}{D_i \cup D_k}, \frac{D_j \cap D_k}{D_j \cup D_k}\right) \tag{8}$$

where $CT_{max} = max_{\forall r,s} CT_{r,s}$ and D_i denotes the set of data points belonging to cluster C_i , α is a constant indicating the confidence level of accepting two non-identical clusters as being similar and its value is set to be 0.8 as default. Based on the weighted G, we can create a refined cluster-association matrix $R \in [0,1]^{n \times r}$. For each ensemble component π_i and their corresponding clusters $\{C_1^i, C_2^i, \dots, C_{k_i}^i\}$ $(i = 1, \dots, M)$, the similarity $R\left(x_j, C_l\right) \in [0,1]$ that data point $x_i \in X^{(i)}$ has with each cluster $C_l \in \{C_1^i, C_2^i, \dots, C_{k_i}^i\}$ (*i*=1, 2, …, *P*) is calculated by

$$R\left(x_j, C_l\right) = \begin{cases} 1, & C_l = C^*(x_i) \\ Sim\left(C_l, C^*(x_i)\right), & otherwise \end{cases} \tag{9}$$

where $C^*(x_i)$ is a cluster label to which data point x_i has been assigned. The refined matrix *R* can be taken as a new representation of the original data, where each column represents the association degree of data points to a specific cluster.

IV. Experimental Results and Discussion

A. Data Set

In our experiments, we present a series of experimental results on real world data to demonstrate the performance of our Stratified Sampling-based Cluster Ensemble algorithm. We select a five cross-language data set for our experiments, the Reuters RCV1/RCV2 Multilingual data set introduced in [17]. This Reuters data set has been used in [18] to evaluate the performance of multi-view spectral clustering algorithm. The data set contains documents are originally written in five different languages, namely English (EN), French (FR), German (GR), Italian (IT) and Spanish (SP). Each document is written in one language originally, and translated to the other four languages using the Portage system [19]. Table I shows us the statistics of the data detail. The documents are categorized into six different topics.

TABLE I: Statistics of the Data Set

Language	Docs	Words
English	18,758	21,531
French	26,648	24,839
German	29,953	34,279
Italian	24,039	15,506
Spanish	12,342	11,547

TABLE II: STATISTICS OF THE DATA SET

Topics	Docs	Percentage
C15	18,816	16.84
CCAT	21,426	19.17
E21	13,701	12.26
ECAT	19,198	17.18
M11	19,421	17.39
GCAT	19,178	17.16

We choose this data set with five views and high dimensions to prove the reliability of SSCE algorithm. And we utilize NMF as our main clustering algorithm. To facilitate the experiments, we randomly selected 1000 documents from each topic. Meanwhile, we remove the unimportant words via Document Frequency (DF), in the experiments, only terms with DF>30 and DF<3000 are kept.

B. Methodology

To demonstrate how the clustering performance can be improved by the proposed approach. We compared with the following algorithms:

- **Single view (BSV** and **WSV):** Run each view by using the same clustering algorithm. We choose the NMF technique as our main clustering algorithm. After get the results of each view, we calculate the performance of each view. We report the best and the worst single view results in our latter sections.
- **Feature Concatenation (ConcatNMF):** Concatenate the features of all the views, and then run NMF directly on this concatenated view representation.
- **Multi-view NMF (MultiNMF):** This is the multi-view NMF clustering algorithm proposed in [9]. The key idea of it is to formulate a joint matrix factorization process with the constraint that pushes clustering solutions of each view towards a common consensus instead of fixing it.
- **Simple Cluster Ensemble (SCE):** This method takes each view as one component and ensembles the component results with link-based confusion function as the final clustering result.
- **Stratified Sampling-based Cluster Ensemble (SSCE-CLC):** This is our proposed method to solve the multi-view clustering problem. This novel method takes the relationship between multiple views into account.

To compare the performance of all the methods, we use four evaluation methods, namely, the accuracy (CA), the rand index (RI), the adjust rand index (AR) and the normalized mutual information (NMI).

CA: It measure the number of correctly classified data points of a clustering solution compared with known class labels. The higher the better.

$$\mathrm{CA} = \frac{\sum_{i=1}^{K}(m_i)}{N} \tag{10}$$

where N is the total number of data in the data set. m_i is the number of data points which correctly categorized to cluster i.

RI: This validity measure takes into account the number of object pairs that exit in the same and different clusters. The RI has a value between 0 and 1, with the more the value approximates to 1 the higher the agreement is.

$$\mathrm{RI} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}\binom{n_{i,j}}{2} + t_3}{\binom{n}{2}} \tag{11}$$

AR: This validity measure is to correct the main criticisms of the Rand Index. Note that the higher the AR value is, the greater the agreement becomes.

$$\mathrm{AR} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}\binom{n_{i,j}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3} \tag{12}$$

where n_i is the number of objects in cluster i, n_j is the number of objects in class j, $n_{i,j}$ is the number of objects occurring in both cluster i and class j, n is the total number of objects in the data set, $t_1 = \sum_{i=1}^{k}\binom{n_i}{2}$, $t_2 = \sum_{j=1}^{k}\binom{n_j}{2}$, $t_3 = \frac{2t_1t_2}{n(n-1)}$.

NMI: The normalized mutual information is to measure the clustering performance. The higher value means better performance.

$$\mathrm{NMI} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{i,j}\log\frac{nn_{i,j}}{n_i n_j}}{\sqrt{\sum_{i=1}^{k} n_i \log\frac{n_i}{n}\sum_{j=1}^{k} n_j \log\frac{n_j}{n}}} \tag{13}$$

C. Experimental Result and Discussion

Table III lists the clustering results of five methods including our proposed SSCE-CLC in terms of four evaluation measures. In order to randomize the experiments, 10 test runs with different data subsets (1000 documents from each topic) were conducted and the average performance are reported.

Obviously, SSCE-CLC consistently outperforms other methods. Especially, it can obtain better result than BSV, i.e., the best single view result. This indicates that the proposed method has ability to make use of the relationship between different views, i.e., different language, to improve the document clustering performance. For ConcatNMF, it just combines the cross-language data in the view of feature level. However, the later NMF clustering algorithm treats the features independently when it identified the clustering coefficients, i.e., the relationship between multiple views is not actually considered. That is the main reason that ConcatNMF is worse than SSCE-CLC.

For MultiNMF, it adopted a joint matrix factorization to find the clustering result V^*. The information from different views simultaneously affects the learning process of V^* (as shown in [9]), that is why it can output better result than ConcatNMF. However, it is sensitive to the parameters and it is hard to select the proper values. In this experiment, we tuning the parameter for each λ_v in range {0.001, 0.01, 0.1, 1, 10, 100, 1000}, and we have to run MultiNMF 16807 ($=7^5$) times to find the best result, which is very time-consuming.

SCE is simple clustering ensemble for cross-language clustering, where the documents set in each language is taken as one component data set of the clustering ensemble. In this case, SCE, like ConcatNMF, totally ignores the relationships between different views.

In order to investigate how the proposed SSCE-CLC works well, we demonstrate the performance improvement of our SSCE-CLC method on each language is shown in Table IV. For each language, two kinds of results are given as two

lines. The first line gives the clustering result with the corresponding single view. The second line lists the improved percentage of SSCE-CLC on the results in the first line.

TABLE III: CLUSTERING PERFORMANCE ON 5-VIEW DATA SET

Method	AR	RI	CA	NMI
BSV	0.2964	0.7958	0.5707	0.3816
WSV	0.2504	0.7764	0.5098	0.3230
ConcatNMF	0.2949	0.7962	0.5368	0.3811
MultiNMF	0.4022	0.7903	0.5237	0.4113
SCE	0.3427	0.8058	0.5917	0.4029
SSCE-CLC	**0.4603**	**0.8368**	**0.6622**	**0.4660**

From Table IV, it can be seen that both German and Spanish languages have relatively higher improvement. For English, the improvement is not very big. This is reasonable because English is popular and the documents in English are easy to be clustered. But for Spanish and German, it is hard to obtain good clustering result after all they belong to minority language. Fortunately, our proposed SSCE-CLC can help improving the clustering performance of German and Spanish documents because it takes advantage of the information from majority language such as English, French and Italian.

TABLE IV: THE PERFORMANCE IMPROMENT PERCENTAGE OF SSCE ON EACH LANGUAGE (%) ON EACH SINGLE VIEW (FOR EACH LANGUAGE, THE FIRST LINE GIVES THE PERFROMANCE ON SINGLE VIEW, THE SECOND LINE GIVES THE IMPROVEMENT)

Language	AR	RI	CA	NMI
English	0.2964 42.61	0.7970 4.43	0.5707 20.34	0.3816 27.44
French	0.2649 59.57	0.7859 5.90	0.5255 30.69	0.3551 36.95
German	0.2734 54.61	0.7958 4.59	0.5102 34.61	0.3267 48.85
Italian	0.2676 57.96	0.7900 5.35	0.5233 31.24	0.3444 41.20
Spanish	0.2504 **68.81**	0.7764 **7.20**	0.5098 **34.72**	0.3230 **50.56**

As we know, the diversity of sampling components is a key component of clustering ensemble [16]. Thus, it is necessary to check the diversity of components generated by our proposed stratified sampling method. Fig. 2 gives the relationship between diversity and quality of component clustering results with our SSCE-CLC method. The horizontal axis is the quality and the vertical axis is the diversity.

Each point was computed from one pair of components clustering created with the stratified sampling component data generation method on data set Reuters. The vertical dashed line in figure indicates the NMI value between the clustering results of an ensemble generated with the LB consensus function from 20 component clustering results and the true labels. There are total 190(C_{20}^2) points in the Fig. 2. The x-axis is the average NMI of the pair of components result computed according to the true labels. The y-axis is the NMI between two component results. From the figures, we can see that our SSCE-CLC method produced good quality component results without sacrificing the diversity.

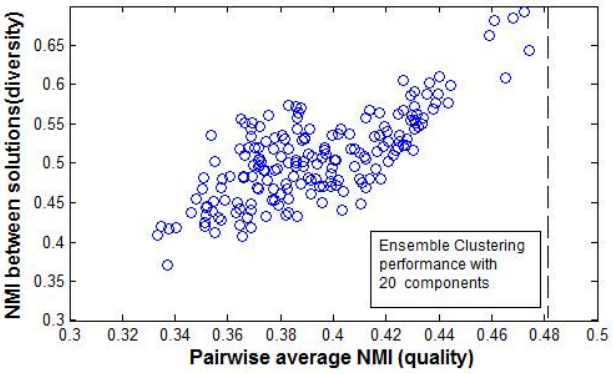

Fig. 2. Demonstration of the relationships between the diversities and qualities of 20 component clustering results which are computed from Reuters data with Stratified Sampling method.

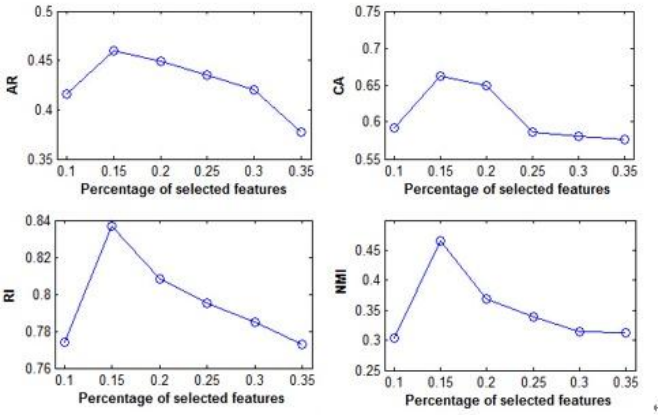

Fig. 3. Effect of sampling size on SSCE-CLC performance.

.

In the experiments, we also investigate the relationship between the performance of clustering ensembles and the sampling rate of features in component data sets. The performance of clustering ensembles of *Reuters* data evaluated in four measures against the sampling rate *p*, i.e., the percentage of features to be sampled in the component data sets is shown in Fig. 3. From this figure, we can see that the performance increased as sampling rate increased, but when sampling increased to 15%, the performance started to drop. It is because the diversity of component clustering results decreases as sampling rate increases to a certain level. We can see that a suitable sampling rate is between 10% and 15% in the Fig. 3. From the results of the experiment on the percentage of clustering ensembles, we set the rate of sampling to 15% for our SSCE-CLC method.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we introduce a novel unsupervised cross-language clustering method, Stratified Sampling-based Cluster Ensemble (SSCE-CLC). In order to make better use of the relationship between multi-view data, our method generates the component data by sampling the features from

all feature groups where each group infers to one language. In this case, the relationships between different languages can be effectively integrated into the clustering process. After generating clustering components, the link-based clustering confusion function is adopted to create the final clustering result, so that the relationship between components can be effectively considered. The proposed SSCE-CLC ensemble method is tested on a real world five cross-language data set. The experimental results have shown that our method has ability to output better performance than other multi-view clustering methods.

In this paper, only the unlabeled data are considered. Even though it is expensive to obtain the label information, there are a few labeled data in real applications. Thus, it is better make use of them in learning process. In the future, we will try semi-supervised multiple view learning under the proposed our stratified sampling framework.

REFERENCES

[1] A. Blum and T. Mitchell, "Combing labeled and unlabeled data with co-training," in *Proc. Annual Conference on Learning Theory (COLT)*, New York, 1998, vol. 4, pp. 570-578.

[2] G. Yuhong, "Convex subspace representation learning from multi-view data," *Association for the Advancement of Artificial Intelligence*, pp. 387-393, vol. ED-11, 2013.

[3] C. Christoudias and U. R, T. Darrell, "Multi-view learning in the presence of view disagreement," in *Proc. Conference on Uncertainty in Artificial Intelligence*, 2008, vol. 10, no. 5, pp. 767-782.

[4] M. Collins and Y. Singer, "Unsupervised models for named entity classification," in *Proc. Conference on Empirical Methods in Natural Language Proceeding (EMNLP)*, 1999, pp. 189-196.

[5] G. Yuhong and X. Min, "Cross language text classification via subspace co-regularized multi-view learning," in *Proc. International Conference on Machine Learning (ICML)*, 2012.

[6] D. Sa, "Spectral clustering with two views," in *Proc. Workshop on Learning with Multiple Views of International Conference on Machine Learning (ICML)*, 2005.

[7] K. Chaudhuri, S. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proc. International Conference on Machine Learning (ICML)*, 2009.

[8] D. Zhou and C. Burges, "Spectral clustering and transductive learning with multiple views," in *Proc. International Conference on Machine Learning (ICML)*, 2007, pp. 1159-1166.

[9] L. Jialu, W. Chi, G. Jing, and H. Jiawei, "Multi-view clustering via Joint Nonnegative Matrix Factorization," in *Proc. SIAM International Conference on Data Mining*, 2013, pp. 252-260.

[10] D. Lee and S. Seung, "Learning the parts of objects by non-negative matrix factorization." *Nature*, 401(6755): 788-791,1999.

[11] A. Singh and G. Gordon, "Relational learning via collective matrix factorization," *KDD*, pp. 650-658, 2008.

[12] D. Freedman, R. Pisani, and R. Purves, *Statistics*, 4th ed., New York: Norton.

[13] T. Liu, F. Wang, and G. Agrawal, "Stratified sampling for data mining on the deep web," in *Proc. ICDM*, pp. 324-333.

[14] S. Fernandes, C. Kamienski, D. Mariz, and D. Sadok, "A stratified traffic sampling methodology for seeing the big picture," *Computer Networks*, vol. 52, pp. 2677-2689, 2008.

[15] Y. Yuming, W. Qingyao, H. J. Zhexue, M. Ng, and L. Xutao, "Stratified sampling for feature subspace in random forests for high dimensional data," *Pattern Recognition*, pp. 769-787.

[16] N. Iam-on and S. Garrett, "LinkCluE: A MATLAB package for link-based cluster ensemble," *Journal of Statistical Software*, vol. 36, issue 9, pp. 1-36, August 2010.

[17] M. R. Amini, N. Usunier, and C. Goutte, "Learning from multiple partially observed views – An application to multilingual text categorization," *Advances in Neural Information Processing Systems (NIPS)*, pp. 28-36, 2009.

[18] A. Kummar, P. Rai, and H. Daume III, "Co-regularied multi-view spectral clustering," *Advances in Neural Information Processing Systems (NIPS)*, pp. 1413-1421, 2011.

[19] N. Ueffing, M. Simard, S. Larkin, and J. H. Johnson, "NRC's PORTAGE system for WMT2007," in *Proc. InACL-2007 Second Workshop on SMT*, 2007, pp. 185-188.

Wenli Gui received the B.S. degree in computer science from Hohai University Wentian College in Maanshan, Anhui, China, in 2012. Currently she is studying for a M.S. degree in computer science in Beijing Jiaotong University.

Since 2013, she was a graduate in Beijing Jiaotong University. She is currently studying in multi-view learning based cross-language mining.

Ms. Gui is now studying for master degree in Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, China.

Liping Jing received the B.S. and M.S. degrees in computer science from Beijing Jiaotong University, Beijing, China, in 2000 and 2003, respectively, and the Ph.D. degree in applied mathematics from the University of Hong Kong, Hong Kong, in 2007.

Since 2007, she was engaged as a research associate in Department of Mathematics, Hong Kong Baptist University, and the Laboratory for Bioinformatics and Medical Informatics in Department of Computer Science, The University of Texas at Dallas, Richardson. She is currently an assistant professor in School of Computer and Information Technology, Beijing Jiaotong University. Her research interests include data mining, text mining, bioinformatics, and business intelligence. She is author of more than 20 peer-reviewed research papers in various journals and conferences.

Dr. Jing has served as a regular reviewer and program committee for a number of international journals and conferences.

Liu Yang received the B.S. and M.S. degrees in computer science from Hebei University, Baoding, Hebei, China, in 2002 and 2006, respectively.

She is a Ph.D. candidate at Beijing Jiaotong University, Beijing, China. Her research interests include transfer learning, image processing, text mining and machine learning.

Dr. Yang is now studying for her Ph.D. degree in Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, China.

Jian Yu received B.S. degree in applied mathematics, the M.S. degree in mathematics, and the Ph.D. degree in applied mathematics from Peking University, Beijing, China, in 1991, 1994, and 2000, respectively.

During 1994-1998, he joined the Faculty of the Beijing Graduate School, China University of Mining and Technology, Beijing. He is currently a professor and the head of the Department of Computer Science, Beijing Jiaotong University, Beijing, China. His current research interests include fuzzy clustering, patten recognition, and data mining.

Prof. Yu has served as a regular reviewer and program committee for a number of international journals and conferences.