Collaborative Multi-View Clustering

Mohamad Ghassany Nistor Grozavu Younès Bennani

Abstract— **The purpose of this article is to introduce a new collaborative multi-view clustering approach based on a probabilistic model. The aim of collaborative clustering is to reveal the common underlying structure of data spread across multiple data sites by applying clustering techniques. The strength of the collaboration between each pair of data repositories is determined by a fixed parameter. Previous works considered deterministic techniques such as Fuzzy C-Means (FCM) and Self-Organizing Maps (SOM). In this paper, we present a new approach for the collaborative clustering using a *generative* model, which is the Generative Topographic Mappings (GTM). Maps representing different sites could collaborate without recourse to the original data, preserving their privacy. We present the approach for *multi-view* collaboration using GTM, where data sets have the same observations but presented in different feature space; i.e. different dimensions. The proposed approach has been validated on several data sets, and experimental results have shown very promising performance.**

I. Introduction

COLLABORATIVE CLUSTERING [1] is an emerging problem in data mining and only some work on this subject have been made in the literature, [1] [2] [3] [4] [5].

In this paper, we assume that we have a group of data sets distributed on different sites; data could be describing customers of banking institutions, stores, medical organizations, etc. The ultimate goal of every organization is to find out some key relationships in its data set. This discovering could be finest by taking into account the dependencies between the different analysis carried out by various sites, in order to produce an accurate view of the global hidden structure in different data sets without sharing data between them.

While most of distributed data clustering (DDC) [6] [7] form a consensus taking into account all their data sets, the fundamental concept of collaboration is that the clustering algorithms operate locally (namely, on individual data sets) but collaborate by exchanging information about their findings [1].

The data sets could include data about different individuals described by the same variables; in this case a vertical collaboration approach is proposed. The horizontal approach for collaboration is used for data sets that describe the same objects but with different variables. This approach can be seen as a multi-view clustering where the treatment is done on multi- represented data, i.e., the same set of objects described by several representations (variables).

Collaborative clustering is divided into two phases: a local phase and a phase of collaboration. The local phase would apply a clustering algorithm based on prototypes, locally and independently on each database. The phase of collaboration aims to collaborate each of the databases with clustering findings associated to other databases obtained from the local phase. Thus, as a result, we obtain on each site a clustering results similar to the results that we would obtain if we had ignored the constraint of confidentiality, i.e. to collaborate databases themselves. At the end of the two phases, all the local clustering will be enriched.

The quality of clustering after the collaboration depends on the chosen method of clustering during the local phase. Previous works on collaborative clustering were based on deterministic models, such as fuzzy c-means (FCM) [1] and self-organizing maps (SOM) [5].

In deterministic models, every set of variable states is uniquely determined by parameters in the model and by sets of previous states of these variables. Therefore, deterministic models perform the same way for a given set of initial conditions. Conversely, in a stochastic model, randomness is present, and variable states are not described by unique values, but rather by probability distributions. In this paper we propose a collaborative multi-view (horizontal) clustering approach based on a probabilistic model, which is the Generative Topographic Mapping (GTM) [8], an alternative topographic clustering to SOM.

The rest of the paper is organized as follows: after introducing the principle of the collaborative clustering in Section 2, we present the principle of the GTM and its EM (Expectation Maximization) algorithm in Section 3. Our proposed Collaborative Multi-View Generative Model is presented in section 4. In Section 5 we present the validation of the proposed approach on different data sets. Finally the paper ends with a conclusion and some future works for the proposed method.

II. Collaborative Clustering

Data Clustering is the main task of knowledge discovery in data sets. Data in this case consists of a set of input vectors without any corresponding target values. The goal in such *unsupervised learning* problem may be to discover groups of similar examples within the data, where it is called *clustering*, or to determine the distribution of data within the input space, known as density estimation, or to project the data from a high-dimensional space down to two or three dimensions for the purpose of *visualization*. Data Clustering aims to group a set of objects in such a way that objects in the same group (called cluster) are more similar to each other than to those in other clusters.

In distributed data clustering (DDC), data are distributed among several sites. The traditional solution to this problem is to collect all the distributed data sets into one centralized

Mohamad Ghassany, Nistor Grozavu and Younès Bennani are with LIPN-UMR 7030, Université Paris 13, Sorbonne Paris Cité, 99, av. J-B Clément, 93430 Villetaneuse, France (email: {firstname.secondname}@lipn.univ-paris13.fr).

repository where the clustering of their union is computed and transmitted back to the sites [6], [7]. This approach, however, may be impractical due to the confidentiality of data. The sites are not be allowed to share data due to legal imposition, e.g. medical records and marketing secrets. DDC techniques aggregate (or fuse) the clustering results into one set to form a consensus, then apply a clustering technique on this consensus taking into account all their data sets, and taking in consideration the confidentiality of data. But in some cases, due to some technical problems, clustering a single large data set may not be feasible. So, a collaborative approach would distribute the classification and merge the different results.

Having distributed data sets on several different sites, the problem is to cluster each of these data sets by considering only the local data and the distant results of clusterings from other data sets (data views), without sharing the data.

The fundamental concept of collaboration is that the clustering algorithms operate locally (namely, on individual data sets) but collaborate by exchanging information about their findings [1] on different views (in the case of horizontal collaboration). So, the data sites exchange information granules during the learning process, taking into consideration the confidentiality of the data. Therefore, in collaborative clustering a phase of collaboration is proposed, aiming to collaborate each of the data views with all clustering results associated to other data views obtained from a local phase applied independently on each data view.

Figure 1 shows the difference between clustering by building a consensus and collaborative clustering.

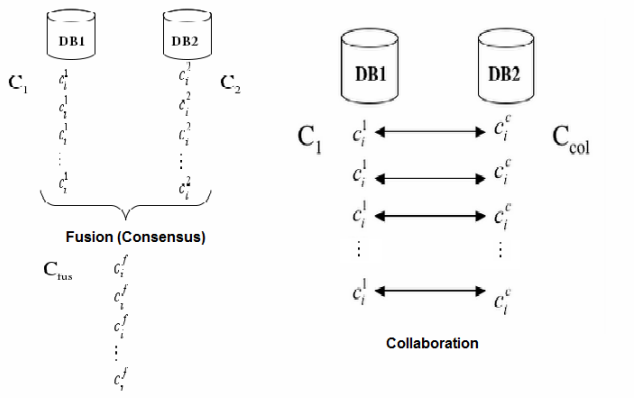

Fig. 1. Building a consensus (left) vs Collaborative Clustering (right)

A topological collaborative clustering (for both horizontal and vertical approaches) was proposed by Grozavu and Bennani [9] inspired from the works of Pedrycz et al. [1] on the c-means collaborative clustering. These approaches are based on the Fuzzy c-means collaborative clustering and introduces the concept of the self-organization firstly introduced by Kohonen [10]. The SOM approaches are often used because they allows clustering and visualization simultaneously for different types of data. This technique can project the data on discrete spaces that are usually in two dimensions. However, the SOM algorithm have some disadvantages:

- Neighborhood-preservation is not guaranteed.
- Convergence of the prototypes is not guaranteed neither.
- There is no theory for initializing the parameters.

III. The Generative Topographic Mapping model

GTM was proposed by Bishop et al. [8], [11] as a probabilistic alternative to SOM [10]. GTM is defined as a mapping from a low dimensional latent space onto the observed data space. The mapping is carried through by a set of basis functions generating a constrained mixture density distribution. It is defined as a generalized linear regression model:

$$y = y(z, W) = W\Phi(z) \tag{1}$$

where y is a prototype vector in the D-dimensional data space, Φ is a matrix consisting of M basis functions $(\phi_1(z), \ldots, \phi_M(z))$, introducing the non-linearity, W is a $D \times M$ matrix of adaptive weights w_{dm} that defines the mapping, and z is a point in latent space. The standard definition of GTM considers spherically symmetric Gaussians as basis functions and is defined as:

$$\phi_m(x) = \exp\left\{-\frac{\|x - \mu_m\|^2}{2\sigma^2}\right\} \tag{2}$$

where μ_m represents the centers of the basis functions and σ - their common width.

Let $\mathcal{D} = (x_1, \ldots, x_N)$ be the data set of N data points. A probability distribution of a data point $x_n \in \Re^D$ is then defined as an isotropic Gaussian noise distribution with a single common inverse β variance:

$$\begin{aligned} p(x_n|z, W, \beta) &= \mathcal{N}(y(z, W), \beta) \\ &= \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{-\frac{\beta}{2}\|x_n - y(z, W)\|^2\right\} \end{aligned} \tag{3}$$

The distribution in x-space, for a given value of W, is then obtained by integration over the z-distribution

$$p(x|W, \beta) = \int p(x|z, W, \beta)p(z)\,dz \tag{4}$$

and this integral can be approximated by defining $p(z)$ as a set of K equally weighted functions on a regular grid,

$$p(z) = \frac{1}{K}\sum_{i=1}^{K}\delta(z - z_k) \tag{5}$$

So, the equation (4) becomes

$$p(x|W, \beta) = \frac{1}{K}\sum_{i=1}^{K}p(x|z_i, W, \beta) \tag{6}$$

For a data set $\mathcal{D}$, we can determine the parameter matrix W, and the inverse variance β, using the maximum

likelihood. In practice it is convenient to maximize the log likelihood, given by:

$$\begin{aligned}\mathcal{L}(W,\beta) &= \ln \prod_{n=1}^{N} p(x_n|W,\beta) \\ &= \sum_{n=1}^{N} \ln \left\{ \frac{1}{K} \sum_{i=1}^{K} p(x_n|z_i,W,\beta) \right\} \end{aligned} \quad (7)$$

The EM Algorithm

The maximization of (7) can be regarded as a missing-data problem in which the identity i of the component which generated each data point x_n is unknown. The EM algorithm for this model is formulated as presented follows.

The posterior probabilities, or responsibilities, of each Gaussian component i for every data point x_n using Bayes' theorem are calculated in the E-step of the algorithm in this form:

$$\begin{aligned} r_{in} &= p(z_i|x_n, W_{old}, \beta_{old}) \\ &= \frac{p(x_n|z_i, W_{old}, \beta_{old})}{\sum_{i'=1}^{K} p(x_n|z_i', W_{old}, \beta_{old})} \\ &= \frac{\exp\{-\frac{\beta}{2}\|x_n - W\phi(z_i)\|^2\}}{\sum_{i'=1}^{K} \exp\{-\frac{\beta}{2}\|x_n - W\phi(z_i')\|^2\}} \end{aligned} \quad (8)$$

As for the M-step, we consider the expectation of the complete-data log likelihood in the form:

$$\mathbf{E}[\mathcal{L}_{comp}(W,\beta)] = \sum_{n=1}^{N}\sum_{i=1}^{K} r_{in} \ln\{p(x_n|z_i,W,\beta)\} \quad (9)$$

The parameters W and β are now estimated maximizing (9), so the weight matrix W is updated according to:

$$\Phi^T G \Phi W_{new}^T = \Phi^T R X \quad (10)$$

where, Φ is the $K \times M$ matrix of basis functions with elements $\Phi_{ij} = \phi_j(z_i)$, R is the $K \times N$ responsibility matrix with elements r_{in}, X is the $N \times D$ matrix containing the data set, and G is a $K \times K$ diagonal matrix with elements

$$g_{ii} = \sum_{n=1}^{N} r_{in} \quad (11)$$

The parameter β is updated according to the expression:

$$\frac{1}{\beta_{new}} = \frac{1}{ND}\sum_{n=1}^{N}\sum_{i=1}^{K} r_{in}\|x_n - W^{new}\phi(z_i)\|^2 \quad (12)$$

In the proposed Collaborative Clustering approach we will use the GTM and EM as a local step, and an adaptation of the GTM to collaborate the distant maps as described in the following section.

IV. Collaborative Multi-View GTM

According to the structure of data sets to collaborate, there is a variety of detailed schemes, two of them are most essential: horizontal (multi-view) and vertical collaboration.

More descriptively, given data sets $X[1], X[2], \ldots, X[P]$ where P denotes their number and $X[ii]$ stands for the $iith$ data set (we adhere to the practice of using square brackets to identify a certain data set). In *horizontal* clustering we have the same objects that are described in *different* feature spaces, $dim(X[1]) \neq dim(X[2]) \neq \ldots \neq dim(X[P])$, while $X[ii] \neq X[jj]$. We note this case as multi-view description, . In other words, these could be the same collection of patients whose records are developed within each medical institution. In collaborative multi-view clustering, the communication platform is based on the partition matrix, posterior probability in case of GTM. As we have the same objects, this type of collaboration makes sense. The confidentiality of data has not been breached: we do not operate on objects but on the resulting granules/prototypes information (fuzzy relations, that is, the posterior probability matrices). As this number is far lower than the number of data, the low granularity of these constructs moves us far from the original data.

Vertical clustering is complementary to horizontal clustering, here the data sets are described in the same feature space but deal with different observations. The vertical collaborative clustering approach using GTM was presented in [12].

In this paper, we propose an approach for multi-view collaboration between several GTMs. Each data set (a view) is clustered through a GTM. To simplify the formalism, the maps built from various data sets will have the same dimensions and the same structure. To collaborate GTMs, we penalize the complete log-likelihood of the M-step, based on [13], considering the term of penalization as a collaboration term, which will penalizes the difference between the posterior probabilities matrices of different data sets, since prototypes don't have same dimensions (they are represented in different feature spaces) in the multi-view collaboration.

The main idea of the multi-view collaboration is that if an observation from the ii-th data set is projected on cluster j in the ii-th map, then that same observation in the jj-th data set will be projected on the same cluster j of the jj-th map or one of its *neighboring* clusters. In other words, clusters that correspond to different maps should capture the same observations.

Here we formulate the underlying optimization problem implied by penalized EM clustering, and derive the proposed algorithm. Assume there are P sets of data with different views, each subset deals with the same patterns, the number of elements in each subset is the same and is equal to N. The collaboration between two subsets is established through an interaction (confidence) coefficient $\alpha_{[ii]}^{[jj]}$ which describes the intensity of the collaboration. In general, $\alpha_{[ii]}^{[jj]}$ is a non-negative value. The higher the value of this parameter is, the stronger the collaboration between the corresponding data sets will be. In this paper, the collaboration is supposed to

be made between two data sets. Suppose that we seek to find the GTM of the data set $[ii]$ collaborating it with the $[jj]$ data set, in the E-step the posterior probabilities are defined as follows:

$$\begin{aligned} r_{in} &= p(z_i|x_n, W_{old}^{[ii]}, \beta_{old}^{[ii]}) \\ &= \frac{p(x_n|z_i, W_{old}^{[ii]}, \beta_{old}^{[ii]})}{\sum_{i'=1}^{K} p(x_n|z_i', W_{old}^{[ii]}, \beta_{old}^{[ii]})} \\ &= \frac{\exp\{-\frac{\beta^{[ii]}}{2}\|x_n - W^{[ii]}\phi^{[ii]}(z_i)\|^2\}}{\sum_{i'=1}^{K}\exp\{-\frac{\beta^{[ii]}}{2}\|x_n - W^{[ii]}\phi^{[ii]}(z_i')\|^2\}} \end{aligned} \quad (13)$$

where $n \in \{1, \ldots, N\}$.
In the M-step, we find $W^{[ii]}$ and $\beta^{[ii]}$ maximizing:

$$\mathcal{L}^{ver}[ii] = \mathbf{E}[\mathcal{L}_{comp}(W^{[ii]}, \beta^{[ii]})] - \alpha_{[ii]}^{[jj]} \sum_{n=1}^{N}\sum_{i=1}^{K} \frac{\beta^{[ii]}}{2}(r_{in}^{[ii]} - r_{in}^{[jj]})^2 \|x_n - W^{[ii]}\phi^{[ii]}(z_i)\|^2 \quad (14)$$

The second term in the above expression makes the clustering based on the *ii*th subset "aware" of other partitions. It is obvious that if the structures in data sets are similar, then the differences between the responsibility matrices tend to be lower, and the resulting structures start becoming more similar.

We derivate (14) w.r.t $W^{[ii]}$ and we put it equal to 0. This leads to write the solution in the following form:

$$\left(\Phi^{[ii]^T} G \Phi^{[ii]} + \alpha_{[ii]}^{[jj]} \Phi^{[ii]^T} F^{[jj]} \Phi^{[ii]}\right) W_{new}^{[ii]^T} = \Phi^{[ii]^T} RX + \alpha_{[ii]}^{[jj]} \Phi^{[ii]^T} H^{[jj]} X \quad (15)$$

where, Φ is the $K \times M$ matrix of basis functions with elements $\Phi_{ij} = \phi_j(z_i)$, R is the $K \times N$ responsibility matrix with elements r_{in}, X is the $N \times D[ii]$ matrix containing the data set, G is a $K \times K$ diagonal matrix, $H^{[jj]}$ is a $K \times N$ matrix, and $F^{[jj]}$ is $K \times K$ diagonal matrix with elements:

$$g_{ii} = \sum_{n=1}^{N} r_{in}^{[ii]} \quad (16)$$

$$h_{in}^{[jj]} = (r_{in}^{[ii]} - r_{in}^{[jj]})^2 \quad (17)$$

$$f_{ii}^{[jj]} = \sum_{n=1}^{N} h_{in}^{[jj]} \quad (18)$$

By derivating (14) w.r.t $\beta^{[ii]}$ and putting it equal to 0, we obtain:

$$\frac{1}{\beta_{new}^{[ii]}} = \frac{1}{ND[ii]} \sum_{n=1}^{N}\sum_{i=1}^{K} (r_{in}^{[ii]} + \alpha_{[ii]}^{[jj]} h_{in}^{[jj]}) \|x_n - W_{new}^{[ii]}\phi^{[ii]}(z_i)\|^2 \quad (19)$$

The proposed Collaborative Multi-View Clustering method is presented in Algorithm 1.

Algorithm 1 The *Collaborative Multi-View GTM algorithm*

Fix the value of $\alpha_{[ii]}^{[jj]}$, a high value means strength collaboration.
Local step:
for $t = 1$ to N_{iter} **do**
For each $BD[ii]$, $ii = 1$ to P :
Build the map using the classical GTM algorithm as described in Section 2.
Collaboration step: For the collaboration of the $[ii]$ map with the $[jj]$ map:
Update the parameters of the $[ii]$-th map using equations 15 and 19.
end for

V. Experimental results

To evaluate our proposed approach we applied our algorithm on several data sets of different sizes and complexity: Waveform, Wisconsin Diagnostic Breast Cancer (wdbc), Glass and Spambase data set [14].

As criteria to validate the approach we used an internal validity index and an external one. External validation is based on previous knowledge about data, i.e real labels. Internal validation is based on the information intrinsic to the data alone. The internal criterion we used is the Davies-Bouldin (DB) index. The external criterion is the map purity (accuracy).

Purity index

The purity index of a map is equal to the average purity of all the clusters of the map. Larger purity values indicate better clustering.

Assuming we have K clusters c_r, $r = 1, \ldots, K$. First, we calculate the purity of each cluster, which is given by:

$$Pu(c_r) = \frac{1}{|c_r|}\ max_i(|c_r^i|)$$

where $|c_k|$ is the total number of data associated to the cluster c_k, $|c_r^i|$ is the number of objects in c_r with class label i. In other words, $Pu(c_r)$ is a fraction of the overall cluster size that the largest class of objects assigned to that cluster represents. Therefore, the overall purity of the clustering solution is obtained as a weighted sum of the individual cluster purities and given as:

$$Purity = \sum_{r=1}^{K} \frac{|c_r|}{N} Pu(c_r) \quad (20)$$

where K is the number of clusters and N is the total number of objects.

Davies-Bouldin index

The Davies-Bouldin (DB) index [15] is an internal validity index aiming to identify sets of clusters that are compact and well separated. It is calculated as follows:

A similarity measure R_{ij} between clusters c_i and c_j is defined basing on a measure of scatter within cluster c_i,

called s_i, and a separation measure between two clusters, called d_{ij}. Then R_{ij} is defined as follows:

$$R_{ij} = \frac{(s_i + s_j)}{d_{ij}}$$

Then, the DB index is defined as:

$$DB_K = \frac{1}{K} \sum_{i=1}^{K} \max_{j:i \neq j} R_{ij} \tag{21}$$

where K denotes the number of clusters.

The DB_K is the average similarity between each cluster $c_i, i = 1, \ldots, K$ and its most similar one. So, smaller value of DB indicates a better clustering solution, thus having minimum possible similarity with the clusters. In order to compute the DB index of the obtained results, we applied a Hierarchical Clustering on the prototypes matrix of the map in order to cluster the map's cells, in this way we obtain a clustering of each data set (before and after the collaboration). We performed several experiments on four data sets from the UCI Repository [14] of machine learning databases.

Data sets

- *waveform data set*: This data set consists of 5000 instances divided into 3 classes (Figure 2). The original data set included 40 variables, 19 are all noise attributes with mean 0 and variance 1. Each class is generated from a combination of 2 of 3 "base" waves.

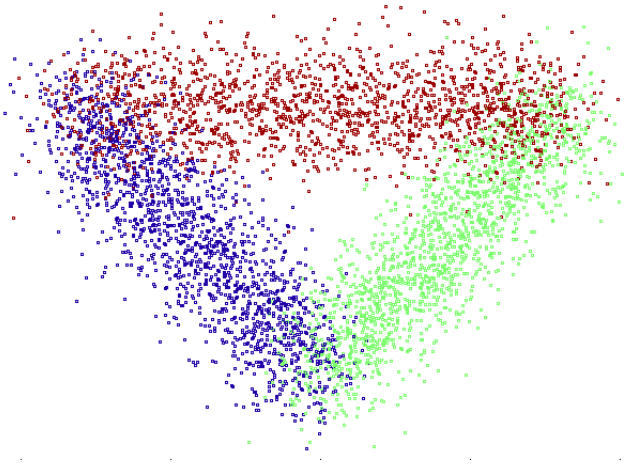

Fig. 2. Waveform original data set, 3 classes of waves are shown.

- *Wisconsin Diagnostic Breast Cancer (WDBC)*: This data has 569 instances with 32 variables (ID, diagnosis, 30 real-valued input variables). Each data observation is labeled as benign (357) or malignant (212).
- *Glass Identification*: Glass Identification data set was generated to help in criminological investigation. At the scene of the crime, the glass left can be used as evidence, but only if it is correctly identified. This data set contains 214 instances , 10 numeric attributes and class name. Each instance has one of 7 possible classes.
- *Spam Base*: The SpamBase data set is composed of 4601 observations described by 57 features. Every feature describes an e-mail and its category: spam or not-spam.

In the following, we will explain the results obtained after applying Collaborative Multi-View GTM algorithm for these data sets. The data sets mentioned above are unified and need to be divided into subsets (or views) in order to have distributed data "scenarios". We divided every data set into two views (subsets) so that the algorithm operates on different features considering, however, the same set of individuals, i.e. Figure 3.

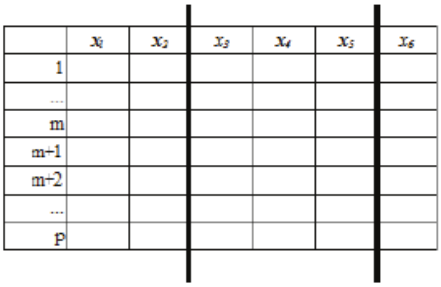

Fig. 3. Horizontal partitioning of data.

First, we applied the local phase, to obtain a GTM map for every subset. We call the resultant maps GTM_1 and GTM_2 respectively for the first and the second subset. The size of all the used maps were fixed to 10×10 except for the Glass data set whose map size is 5×5. Then we applied the collaboration phase, in which we seek a new GTM for the subset but collaborating it with the other subset. We call $GTM_{2\to1}$ the map representing subset 1 and receiving information (clustering results) from subset 2.

As described above, the waveform data set is composed from two subsets of variables: the variables from 1 to 21 representing relevant characteristics, variables from 22 to 40 are noisy. This data structure allows us to divide the data set in two views: first one containing relevant variables and the second one containing only the noisy variables. Results of local phases using GTM for these two views are presented in Figures 4 and 5 respectively for first and second view. These figures were obtained by projecting the data into two dimensional space using Principal Component Analysis [16], [17] applied on the waveform data set, but the color of the points represent the class of each object obtained using GTM and followed by a majority vote rule on the first subset (Figure 4) and on the second noisy set respectively (Figure 5).

Note, that for a better understanding of the results, the figures should be analyzed in a color mode.

The three classes of the waveform data set are well represented and separated on Figure 4. While they are not in Figure 5 due to the variables noisiness of this view.

After applying the collaboration to exchange the clustering information between all the maps without sharing data between them, we obtained the following:

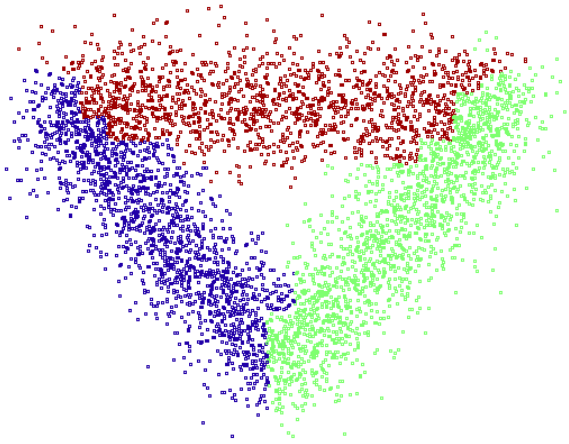

Fig. 4. waveform subset 1, relevant variables: labeling data using GTM_1

Fig. 5. waveform subset 2, noisy variables: labeling data using GTM_2

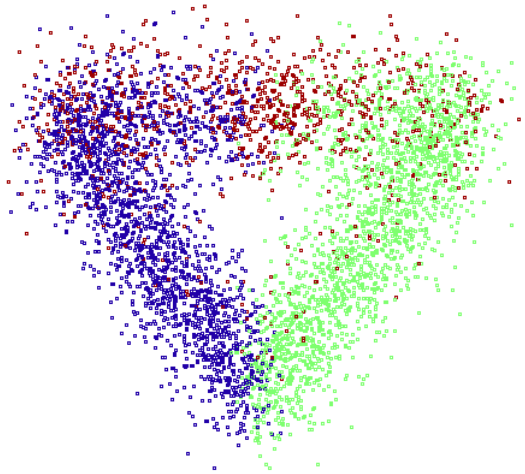

Fig. 6. waveform subset 1 after collaboration with subset 2: labeling data using $GTM_{2\to1}$

Fig. 7. waveform subset 2 after collaboration with subset 1: labeling data using $GTM_{1\to2}$

After the collaboration of the first view (relevant variables of the waveform data set) with the noisy variables clustered by GTM_2 map, the purity index decreased from 86.25% to 72.78% (Table I). Figure 7 shows the projection of data by labeling them using the results of the collaborated map $GTM_{2\to1}$, we can see that clusters are not well separated comparing to what we have obtain before collaboration in Figure 4.

Contrarily, by applying the collaboration in the opposite direction, the purity index of the $GTM_{1\to2}$ map increased from 38.47% to 57.12% compared to the GTM_2. Results are shown in Figure 6 in which we can see that clusters are better separated now after collaboration of noisy variables with relevant variables.

Results explained above are reasonable and show the importance of collaboration. When a clustering of a set described by relevant variables collaborate with a clustering of a set containing noisy variables, the quality of clustering decreases. While in the opposite case, sending clustering results of a set described by relevant variables to the clustering of noisy set increases its quality.

As for the other data sets, we divided them all to two views. We computed the purity index and the DB index before and after collaboration and the results are shown in Table I.

In most of the cases, we remark that the purity of the map is getting higher or do not change drastically after the collaboration and strongly depends on the relevance of the collaborative map (the quality of the collaborative classification). The same analysis can be made for the DB index which decreases after the collaboration using a relevant map. For example the DB index of the $GTM_{2\to1}$ for Glass data set obtained using the information from GTM_2 during the learning of the GTM_1 decreases from 1.28 to 0.97 (Table I). This shows an amelioration of the clustering results.

This conclusion corresponds to the intuitive understanding of the principle and to the consequences of such cooperation. However, note that the goal was not to improve the clustering

TABLE I

EXPERIMENTAL RESULTS OF THE MULTI-VIEW COLLABORATIVE APPROACH ON DIFFERENT DATA SETS

Dataset	Map	Purity (%)	DB Index
Waveform			
4000x21	GTM_1	86.25	1.14
4000x19	GTM_2	38.47	3.75
	$GTM_{1\to2}$	57.12	1.73
	$GTM_{2\to1}$	72.78	1.31
Glass			
214x5	GTM_1	92.32	0.74
214x5	GTM_2	64.02	1.28
	$GTM_{1\to2}$	73.42	1.05
	$GTM_{2\to1}$	83.18	0.97
Wdbc			
569x16	GTM_1	94.07	0.97
569x16	GTM_2	96.27	0.87
	$GTM_{1\to2}$	95.88	0.9
	$GTM_{2\to1}$	94.92	0.92
SpamBase			
4601x28	GTM_1	80.17	1.12
4601x28	GTM_2	84.26	0.95
	$GTM_{1\to2}$	83.35	0.98
	$GTM_{2\to1}$	82.61	1.06

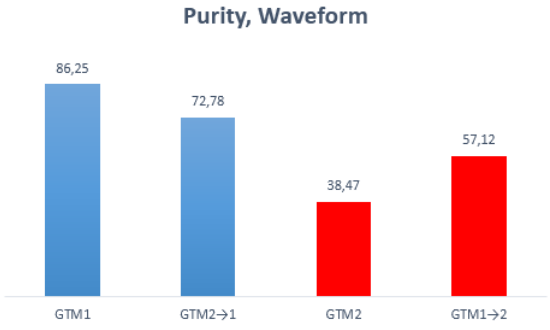

Fig. 8. Comparison of the purity obtained for Waveform subsets, before and after the collaboration

accuracy but to take into account the distant information and to build a new map using another view of the same data, and this procedure can decrease sometimes the quality of clustering which depends on the variables relevance of the view to collaborate.

After doing the collaboration by applying the proposed method, a next important step is to precise how the collaboration can improve all the results, i.e., how every clustering helps to improve the overall clustering, accepting good clustering and rejecting bad clustering. This will be done by estimating the coefficient of collaboration $\alpha_{[ii]}^{[jj]}$ during the learning process. An example is shown in [3], [5] and [4]. This coefficient will precise the confidence that data set $[ii]$ gives to data set $[jj]$, i.e., a value indicating how much data set $[ii]$ trusts data set $[jj]$, more it is high more the confidence level is high. Depending on these estimated values, we can form a consensus clustering. Here is the true potential of Collaborative Clustering and why it is better than standard distributed data clustering algorthims. It is like forcing the data sets to discuss between them before forming the consensus instead of forming it directly without taking into account whether the results of clustering are good or bad.

VI. CONCLUSION

In this study we proposed a methodology to apply a Collaborative Multi-View Clustering on distributed data. The proposed algorithm is based on GTM as a local phase of clustering, and an extension of it in the collaboration step. The Collaborative Multi-View learning approach is adapted to the problem of collaboration of several data sets containing the same observations described different variables. During the collaboration step, we do not need the share the data between sites but only the results of the distant clustering. Thus, each site uses its data view and the information from other clusterings, which would provide a new clustering that is as close as possible to that which would be obtained if we had centralized the data sets. We presented and approach basing on probabilistic model to cluster the data, which is the Generative Topographic Mappings. We presented the formalism of Collaborative Clustering using an adapted extension of this method. The approach has been validated on multiple data sets and the experimental results have shown promising performance.

Several perspectives can be considered for this work as: to add a step in the collaboration phase to estimate the value of the coefficients of collaboration; to merge all the clustering results obtained after the collaboration and to build a consensus for all the data views.

REFERENCES

[1] W. Pedrycz, "Collaborative fuzzy clustering," *Pattern Recognition Letters*, vol. 23, no. 14, pp. 1675–1686, 2002.

[2] W. Pedrycz and K. Hirota, "A consensus-driven fuzzy clustering," *Pattern Recogn. Lett.*, vol. 29, no. 9, pp. 1333–1343, 2008.

[3] N. Grozavu, M. Ghassany, and Y. Bennani, "Learning confidence exchange in collaborative clustering," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, 31 2011-aug. 5 2011, pp. 872–879.

[4] B. Depaire, R. Falcón, K. Vanhoof, and G. Wets, "Pso driven collaborative clustering: A clustering algorithm for ubiquitous environments," *Intell. Data Anal.*, vol. 15, no. 1, pp. 49–68, Jan. 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1937721.1937725

[5] M. Ghassany, N. Grozavu, and Y. Bennani, "Collaborative clustering using prototype-based techniques," *International Journal of Computational Intelligence and Applications*, vol. 11, no. 03, p. 1250017, 2012. [Online]. Available: http://www.worldscientific.com/doi/abs/10.1142/S1469026812500174

[6] A. Strehl, J. Ghosh, and C. Cardie, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2002.

[7] J. C. da Silva and M. Klusch, "Inference in distributed data clustering," *Eng. Appl. Artif. Intell.*, vol. 19, no. 4, pp. 363–369, Jun. 2006. [Online]. Available: http://dx.doi.org/10.1016/j.engappai.2006.01.013

[8] C. M. Bishop, M. Svensén, and C. K. I.Williams, "Gtm: The generative topographic mapping," *Neural Comput*, vol. 10, no. 1, pp. 215–234, 1998.

[9] N. Grozavu and Y. Bennani, "Topological collaborative clustering," *in Australian Journal of Intelligent Information Processing Systems (AJIIPS)*, vol. 12, no. 3, 2010.

[10] T. Kohonen, *Self-organizing Maps*. Berlin: Springer-Verlag Berlin, 1995.

[11] C. M. Bishop and C. K. I. Williams, "Gtm: A principled alternative to the self-organizing map," in *In Advances in Neural Information Processing Systems*. Springer-Verlag, 1997, pp. 354–360.

[12] M. Ghassany, N. Grozavu, and Y. Bennani, "Collaborative generative topographic mapping," in *Neural Information Processing*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7664, pp. 591–598.

[13] P. J. Green, "On Use of the EM Algorithm for Penalized Likelihood Estimation," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 52, no. 3, pp. 443–452, 1990. [Online]. Available: http://dx.doi.org/10.2307/2345668

[14] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[15] D. B. D.L. Davies, "A cluster separation measure," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 1 (4), pp. 224–227, 1974.

[16] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 6, pp. 559–572, 1901.

[17] H. Hotelling, "Analysis of complex statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, Sep. 1933.