Collaborative Generative Topographic Mapping

Mohamad Ghassany, Nistor Grozavu, and Younès Bennani

Université Paris 13, Sorbonne Paris Cité.
LIPN UMR CNRS 7030
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse
FirstName.LastName@lipn.univ-paris13.fr

Abstract. The aim of collaborative clustering is to reveal the common structure of data distributed on different sites. In this paper, we present a new approach for the topological collaborative clustering using a generative model, which is the Generative Topographic Mappings (GTM). In this case, maps representing different sites could collaborate without recourse to the original data, preserving their privacy. Depending ont the data structure, there are three different ways of collaborative clustering: horizontal, vertical and hybrid. In this study we introduce the Collaborative GTM for the vertical collaboration. The article presents the formalism of the approach and its validation. The proposed approach has been validated on several datasets and experimental results have shown very promising performance.

Keywords: Collaborative clustering, Prototype based clustering, Generative Topographic Mapping, Privacy preserving.

1 Introduction

The Collaborative Clustering [7] is an emerging problem in data mining and only some work on this subject have been made in the literature, [7] [8] [4] [5] [2]. In this study, we assume that we have a group of datasets distributed on different sites which could include data about different individuals described by the same variables or the datasets could represent the same samples but with different descriptors (variables). For example, the data could be describing cutomers of banking institutions, stores, medical organizations, etc. The ultimate goal of every organization is to find out some key relationships in its dataset. 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 datasets without sharing data between them. 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 [7]. So we propose an Collaborative Generative Model divided into two phases: a local phase and a phase of collaboration. The local phase would apply a clustering algorithm based on prototypes (classical GTM), locally and independently on each database. The phase

T. Huang et al. (Eds.): ICONIP 2012, Part II, LNCS 7664, pp. 591–598, 2012.

of collaboration would work to collaborate each of the databases with all classifications 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 condentiality, i.e. to collaborate databases themselves. At the end of the two phases, all the local clustering will be enriched.

The rest of this paper is organized as follows: we present the principle of the GTM and EM algorithm in Section 2. Our proposed Vertical Collaborative Generative Model is presented in section 3. In Sections 4, we present the valdiation of the proposed approach on different datasets. Finally the paper ends with a conclusion and some future works for the proposed methods.

2 The GTM Model as a Local Step for the Collaborative Clustering

GTM was proposed by Bishop et al. [1] as a probabilistic counterpart to the Self-organizing maps (SOM) [6]. 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:

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

where y is a prototype vector in the D-dimensional data space, $\varPhi$ 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, 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 variance β:

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

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 defining $p(z)$ as a set of K equally weighted delta functions on a regular grid,

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

So, equation (4) becomes

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

For the data set $\mathcal{D}$, we can determine the parameter matrix W, and the inverse variance β, using maximum likelihood. In practice it is convenient to maximize the log likelihood, given by:

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

2.1 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 follows:

The posterior probabilites, or responsibilites, 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

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

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

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

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

$$\Phi^T G \Phi W_{new}^T = \Phi^T R X \tag{10}$$

where, Φ is the $K \times M$ matrix of basis functions with elements $\Phi_{ij} = \phi_j(z_i)$, R is the $K \times N$ responsability 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

$$g_{ii} = \sum_{n=1}^{N} r_{in} \tag{11}$$

The parameter β is updated according to

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

In the proposed Collaborative Clustering Model we will use the GTM and EM as a local step, and an adaptaion of the GTM to transfer the knowledge from a dataset to a map as described in the following section.

3 Collaborative Generative Topographic Mappings

According to the structure of datasets to collaborate, there are three main types of collaboration principle: horizontal, vertical and hybrid. In this paper, we are specifically interested in vertical collaboration. The vertical collaboration is to collaborate the clustering results obtained from different datasets described by the same variables, but having different objects. In this paper, we study the collaboration between several clus- tering results, especially the collaboration between several Generative topographic mappings. Each dataset is clustered through a GTM, and to simplify the formalism, the maps built from various datasets will have the same dimensions and the same structure.

So, to collaborate GTMs, we will base on [3], considering the term of penalization as a collaboration term, which will penalize the distance between the prototypes of different datasets.

In the vertical collaboration case, all datasets have the same variables (same description space), but have different observations, $N[ii] \neq N[jj]$. In this case, the observations of these datasets have the same size, and the dimension of the the prototype vectors for all the GTMs will be the same. Suppose that we seek to find the GTM of the dataset $[ii]$ collaborating it with the $[jj]$ dataset, the E-step stays as it is, in which we find the posterior probabilities:

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

where $n \in \{1, \ldots, N[ii]\}$.

In the M-step, we find $W^{[ii]}$ and $\beta^{[ii]}$ maximizing

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

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

$$\Phi^{[ii]^T} \left(G\Phi^{[ii]} + \alpha_{[ii]}^{[jj]} G\Phi^{[ii]} \right) W_{new}^{[ii]^T} = \Phi^{[ii]^T} RX - \alpha_{[ii]}^{[jj]} \Phi^{[ii]^T} G\Phi^{[jj]} W^{[jj]^T} \tag{15}$$

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

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

Then,

$$W_{new}^{[ii]^T} = \left(\Phi^{[ii]^T} \left(G\Phi^{[ii]} + \alpha_{[ii]}^{[jj]} G\Phi^{[ii]} \right) \right)^{-1} \left(\Phi^{[ii]^T} RX - \alpha_{[ii]}^{[jj]} \Phi^{[ii]^T} G\Phi^{[jj]} W^{[jj]^T} \right) \tag{17}$$

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

$$\frac{1}{\beta_{new}^{[ii]}} = \frac{1}{N[ii]D} \sum_{n=1}^{N[ii]} \sum_{i=1}^{K} \left[r_{in} \| x_n - W_{new}^{[ii]} \phi^{[ii]}(z_i) \|^2 - \alpha_{[ii]}^{[jj]} r_{in} \| W_{new}^{[ii]} \phi^{[ii]}(z_i) - W^{[jj]} \phi^{[jj]}(z_i) \|^2 \right] \tag{18}$$

Therefore, the proposed Collaborative Clustering method is presented as following:

Algorithme 1. The vertical collaboration GTM algorithm

```
Fix the value of $\alpha_{[ii]}^{[jj]}$, a high value means strengthful 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 prototypes of the $[ii]$ map by using the function 19
**end for**
```

4 Experimental Results

To evaluate our proposed collaborative approache we applied our algorithm on several datasets of different size and complexity. The used datasets are the following: Waveform, Wisconsin Diagnostic Breast Cancer (wdbc), Isolet and Spambase.

As criterion to validate our approach we used the purity (accuracy) index of the map which is equal to the average purity of all the cells of the map. A good GTM map should have a high degree of the purity index.

The purity of cells is the percentage of data belonging to the majority class. Assuming that the data labels set $L = l_1, l_2, ..., l_{|L|}$ and the prototypes set $C = c_1, c_2, ..., c_{|C|}$ are known, the formula that expresses the purity of a map is the following:

$$purity = \sum_{k=1}^{|C|} \frac{c_k}{N} \times \frac{max_{i=1}^{|L|} |c_{ik}|}{|c_k|} \tag{19}$$

where $|c_k|$ is the total number of data associated with the cell c_k, and $|c_{ik}|$ is the number of data of class l_i which are associated to the cell c_k and N - the total number of data.

4.1 Data Sets

- *waveform dataset*: This data set consists of 5000 instances divided into 3 classes. The original base 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.
- *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). Variables are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.
- *Isolet*: This data set was generated as follows. 150 subjects spoke the name of each letter of the alphabet twice. Hence, we have 52 training examples from each speaker. The speakers are grouped into sets of 30 speakers each, and are referred to as isolet1, isolet2, isolet3, isolet4, and isolet5. The data consists of 1559 instances and 617 variables. All variables are continuous, real-valued variables scaled into the range -1.0 to 1.0.
- *Spam Base*: The SpamBase data set is composed from 4601 observations described by 57 variables. Every variable described an e-mail and its category: spam or not-spam. Most of the attributes indicate whether a particular word or character was frequently occurring in the e-mail. The run-length attributes (55-57) measure the length of sequences of consecutive capital letters.

In the following, we will explain the results obtained after applying the Collaborative GTM algorithm for these datasets. The data sets mentioned above are

Table 1. Experimental results of the vertical collaborative approach on different datasets

Dataset	Map	Purity	Dataset	Map	Purity
Waveform	GTM_1	86.44	Isolet	GTM_1	87.17
	GTM_2	86.52		GTM_2	86.83
	$GTM_{1\to 2}$	87.16		$GTM_{1\to 2}$	87.29
	$GTM_{2\to 1}$	87.72		$GTM_{2\to 1}$	85.87
Wdbc	GTM_1	96	SpamBase	GTM_1	52.05
	GTM_2	96.34		GTM_2	51.68
	$GTM_{1\to 2}$	96.08		$GTM_{1\to 2}$	52.41
	$GTM_{2\to 1}$	96.15		$GTM_{2\to 1}$	52.17

unified and need to be divided in subsets in order to have distributed data scenarios. So, we divide every data set into two subsets, having the same features, but with different observations. First, we applied the local phase, to obtain a GTM map for every subset. The size of all the used maps were fixed to 10×10 except for the Isolet dataset whose map size is 5×5. Then we started the collaboration phase, in which we seek a new GTM for the subset but collaborating it with the other subset. We calculated the purity index of the new GTMs after collaboration, results are shown in Table 1.

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) and on the confidence on this map (the collaboration parameter). However, for the Isolet dataset, collaborating the second map with the first subset deacrese the accuracy index from 86.84% to 85.87%. For the SpamBase dataset we have only a small improvement of the purity index when collaborating the both maps. This conclusion corresponds to the intuitive understanding of the principle and to the consequences of such cooperation. Also, note that the goal was not to improve the clustering accuracy but to take into account the distant information and to build a new map using another dataset, and this procedure can deacrease sometimes the acuracy index which depends on the quality of the dataset to collaborate.

5 Conclusion

In this study we proposed a methodology to apply a vertical collaborative 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 phase. The vertical collaborative learning approach is adapted to the problem of collaboration of several datasets containing the same variables but with different observations. During the collaboration step, we do not need the datasets but only the results of the distant classifications. Thus, each site uses its dataset and the information from other classifications, which would provide a new classification that is as close as possible to that which would be obtained if we had centralized the datasets. The approach has been validated on multiple databases and the experimental results have shown promising performance.

Several perspectives can be considered for this work as: to propose an approach for the horizontal case of collaboration; add a step in the collaboration phase to estimate the best values of the coefficients of collaboration; to fuse all the classifications obtained after the collaboration and to build a consensus classification for all the distant sites.

References

1. Bishop, C.M., Svensén, M., Williams, C.K.I.: Gtm: The generative topographic mapping. Neural Comput. 10(1), 215–234 (1998)
2. Depaire, B., Falcon, R., Vanhoof, K., Wets, G.: Pso driven collaborative clustering: A clustering algorithm for ubiquitous environments. Intell. Data Anal. 15(1), 49–68 (2011)
3. Peter, J.G.: On Use of the EM Algorithm for Penalized Likelihood Estimation. Journal of the Royal Statistical Society. Series B (Methodological) 52(3), 443–452 (1990)
4. Grozavu, N., Ghassany, M., Bennani, Y.: Learning confidence exchange in collaborative clustering. In: The 2011 International Joint Conference on Neural Networks (IJCNN), vol. 31, pp. 872–879 (2011)
5. Bennani, Y., Grozavu, N.: Topological collaborative clustering. In: LNCS Springer of ICONIP 2010: 17th International Conference on Neural Information Processing (2010)
6. Kohonen, T.: Self-organizing Maps. Springer, Berlin (1995)
7. Pedrycz, W.: Collaborative fuzzy clustering. Pattern Recognition Letters 23(14), 1675–1686 (2002)
8. Pedrycz, W., Hirota, K.: A consensus-driven fuzzy clustering. Pattern Recogn. Lett. 29(9), 1333–1343 (2008)