International Journal of Computational Intelligence and Applications
Vol. 11, No. 3 (2012) 1250017 (17 pages)

DOI: 10.1142/S1469026812500174

Imperial College Press
www.icpress.co.uk

COLLABORATIVE CLUSTERING USING PROTOTYPE-BASED TECHNIQUES

MOHAMAD GHASSANY*, NISTOR GROZAVU† and YOUNES BENNANI‡

Université Paris 13, Sorbonne Paris Cité, LIPN UMR CNRS 7030
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse, France
**mohamad.ghassany@lipn.univ-paris13.fr*
†*nistor.grozavu@lipn.univ-paris13.fr*
‡*younes.bennani@lipn.univ-paris13.fr*

Received 22 March 2012
Revised 14 June 2012
Published 30 September 2012

The aim of collaborative clustering is to reveal the common structure of data distributed on different sites. In this paper, we present a formalism of topological collaborative clustering using prototype-based clustering techniques; in particular we formulate our approach using Kohonen's Self-Organizing Maps. Maps representing different sites could collaborate without recourse to the original data, preserving their privacy. We present two different approaches of collaborative clustering: horizontal and vertical. The strength of collaboration (confidence exchange) between each pair of datasets is determined by a parameter, we call coefficient of collaboration, to be estimated iteratively during the collaboration phase using a gradient-based optimization, for both the approaches. The proposed approaches have been validated on several datasets and experimental results have shown very promising performance.

Keywords: Collaborative clustering; distributed data; prototype-based clustering; self-organizing maps; privacy preserving.

1. Introduction

In this study, we assume that we have a group of datasets distributed on different sites; data could be describing customers of banking institutions, stores, medical organizations, etc. The datasets could include data about different individuals described by the same variables; in this case we present the vertical collaboration approach. Otherwise, the datasets could represent the same individuals but with different descriptors (variables), reflecting the activities of the organization. Such case is considered to be the most difficult one, since different variables mean that samples are described in different feature spaces, thus different dimension. Therefore we present a horizontal collaboration approach for this case. 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 analyses carried out by various sites. By that, it produces an

accurate view of the global hidden structure in different datasets without sharing data between them. So, fusing all datasets in one large dataset then applying a clustering technique on this large dataset is not feasible because of data confidentiality.

Most of distributed data clustering (DDC)[1,2] 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 datasets, taking in consideration the confidentiality of data. But in some cases, due to some technical problems, the classification of a single large dataset may not be feasible. So, a collaborative approach would distribute the classification and merge the different results.

The fundamental concept of collaboration is that the clustering algorithms operate locally (namely, on individual datasets) but collaborate by exchanging information about their findings.[3] So we propose an approach 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, which will result in obtaining a score for each of these bases. The phase of collaboration would work to collaborate each of the databases with all classifications associated to other databases obtained from the local phase. We consider collaboration is done by pairs. In this paper, we divide "the collaboration phase" into two steps, the first step in which we compute the prototypes matrix, and the second step specified for learning the confidence links.[4] As a result, we obtain on each site 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 Collaborative Clustering[3,5–8] is an emerging problem in data mining and only some work on this subject have been made in the literature. We propose two approaches, one for the horizontal collaborative clustering and one for the vertical collaborative clustering. The horizontal approach for collaboration is used for datasets 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). The vertical approach is for collaborating several datasets of different objects but described by same variables. During the collaboration phase, we do not need the datasets; we only need the results of 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 what would be obtained if we had centralized the datasets and then made a clustering. The rest of this paper is organized as following: we present the principle of collaborative classification in Sec. 3 after a short introduction of Self-Organizing Maps (SOM) algorithm in Sec. 2. Our proposed methodology for collaborative approaches (vertical and horizontal) is presented in Secs. 3.1 and 3.2. The experimental results are presented in Sec. 4 and a conclusion with some future works for the proposed method is presented in Sec. 5.

2. Prototype-Based Clustering Techniques

A large variety of methods of clustering has been developed. Several of these methods are based on very simple fundamentals, yet very effective idea, namely describing the data under consideration by a set of prototypes, which capture characteristics of the data distribution (like location, size, and shape), and to classify or divide the dataset based on the similarity of the data points to these prototypes. The approaches relying on this idea differ mainly in the way in which prototypes are described and how they are updated during the model construction step.

In this paper, we are specifically interested in the topological collaborative clustering (horizontal and vertical collaboration) approaches, proposed by Grozavu and Bennani,[5] and inspired from the works of Pedrycz *et al.*[6,7] on the fuzzy c-means collaborative clustering. These two approaches are based on the Fuzzy c-means collaborative clustering and introduce the concept of the self-organization first introduced in the SOM of Kohonen.[9] The goal of SOM is to find a set of centroids (reference vectors) and to assign each object in the dataset to be the centroid that provides the best approximation of that object. The SOM models are often used because they allow clustering and visualization simultaneously for different types of data. Indeed, this technique can project the data on discrete spaces that are usually in two dimensions. The topological collaborative clustering results straightly depend on the collaboration/confidence matrix. The confidence matrix precise the strength of collaboration, so its choice is critical since setting in advance the strength of the collaboration for each collaboration link (collaboration confidence parameters) can degrade the final results if it is not set correctly. In an unsupervised collaborative learning, no knowledge is available and usually this parameter is set to 1 to avoid unconformity to the collaborative dataset. In this paper, we estimate the values of the collaboration confidence parameters iteratively during the learning process, inspired by Ref. 10.

2.1. *SOM as local step for the collaborative clustering*

The SOM introduced by Kohonen[9] have been widely used for unsupervised classification and visualization of multidimensional datasets. There is a wide variety of algorithms for topological maps derived from the original model proposed first by Kohonen.[11–13] These models are different from each other, but share the same idea to present the large data in a simple geometric relationship on a reduced topology.

The model consists in the attempting of unsupervised classification of a learning set $A = \{x^{(i)} \in \mathbb{R}^n, i = 1, \ldots, N\}$ where $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_j^{(i)}, \ldots, x_n^{(i)})$. This classical model consists in a discrete set C of cells (neurons) called map. This map has a discrete topology defined by undirected graph; usually it is a regular grid in two dimensions. The influence notion of a cell k on a cell l, which depends on their proximity, is presented by a kernel function K ($K \geq 0$ and $\lim_{|x|\to\infty} K(x) = 0$).

The mutual influence between two units k and l is defined by the function $K_{k,l}(.)$:

$$K_{ij} = \frac{1}{\lambda(t)} \exp\left(-\frac{d_1^2(i,j)}{\lambda^2(t)}\right), \tag{1}$$

where $\lambda(t)$ is the temperature's function modeling the neighborhood's range:

$$\lambda(t) = \lambda_i \left(\frac{\lambda_f}{\lambda_i}\right)^{\frac{t}{t_{\max}}} \tag{2}$$

with λ_i and λ_f are the initial temperature and the final temperature (for example $\lambda_i = 2$ and $\lambda_f = 0.5$) and $t_{\max}$ is the maximum allotted time (number of iterations). The Manhattan distance $d_1(.,.)$ between two map units r and s of coordinates (k,m) and (i,j), is defined by:

$$d_1(r,s) = |i-k| + |j-m|. \tag{3}$$

The function $K_{k,l}(.)$ is a Gaussian introduced for each neuron of the map with a global neighborhood. The size of this neighborhood is limited by the standard Gaussian deviation $\lambda(t)$. The units that are beyond this range have a significant influence (but not null) on the considered cell. The range $\lambda(t)$ is a decreasing function with time, so, the neighborhood function $K_{k,l}(.)$ will have the same trend with a standard deviation decreasing in time.

For each cell k of the grid is associated a reference (prototype) vector $w^{(k)} = (w_1^{(k)}, w_2^{(k)}, \ldots, w_i^{(k)}, \ldots, w_n^{(k)})$ of size n. We note by W the set of referents. The learning of this model will be reached by minimizing the distance between input pattern and prototypes of the map, weighted by the neighborhood. A gradient algorithm can be used for this purpose. The criterion to minimize in this case is:

$$R(\chi, W) = \sum_{i=1}^{N} \sum_{j=1}^{C} K_{j,\chi(x^{(i)})} \|x^{(i)} - w^{(j)}\|^2, \tag{4}$$

where χ assigns each pattern (observation) $x^{(i)}$ to a single cell of the SOM.

At the end of the learning, the SOM determines a data partition in C groups associated with each cell k of the map. Each group or cell is associated with a reference vector $w^{(k)} \in \mathbb{R}^n$, which will be the representative, the "local mean" or the prototype of the observation's set associated with this cell.

3. Collaborative Clustering

While collaboration can include a variety of detailed schemes, two of them are the most essential. We refer to them as horizontal and vertical modes of collaboration or simply horizontal and vertical clustering. More descriptively, given datasets $X[1]$, $X[2], \ldots, X[P]$ where P denotes their number and $X[ii]$ stands for the $iith$ dataset (we adhere to the practice of using square brackets to identify a certain dataset), in *horizontal* clustering we have the same objects that are described in *different* feature

spaces. In other words, these could be the same collection of patients whose records are developed within each medical institution. In horizontal clustering we deal with the same patterns and different feature spaces. The communication platform is based on through the partition matrix (Kernels in case of SOM). 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 individual patterns but on the resulting information granules (fuzzy relations, that is, partition 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 datasets are described in the same feature space but deal with *different* patterns. In other words, we consider that $X[1], X[2], \ldots, X[P]$ are defined in the same feature space, while each of them consists of different patterns, $\dim(X[1]) = dim(X[2]) = \cdots = \dim(X[P])$, while $X[ii] \neq X[jj]$. In vertical clustering we are concerned with different patterns but the same feature space. Hence communication at the level of the prototypes (which are high-level representatives of the data) becomes feasible. Again, because of the aggregate nature of the prototypes, the confidentiality requirement has been satisfied. There are also many hybrid models of collaboration involving datasets with possible links of vertical and horizontal collaboration, that are not discussed in the paper.

3.1. *Topological horizontal collaboration*

Here we formulate the underlying optimization problem implied by objective function-based clustering, and derive the detailed algorithm. There are P sets of data located in different spaces (viz., the patterns are described by different features). As each subset deals with the same patterns, the number of elements in each subset is

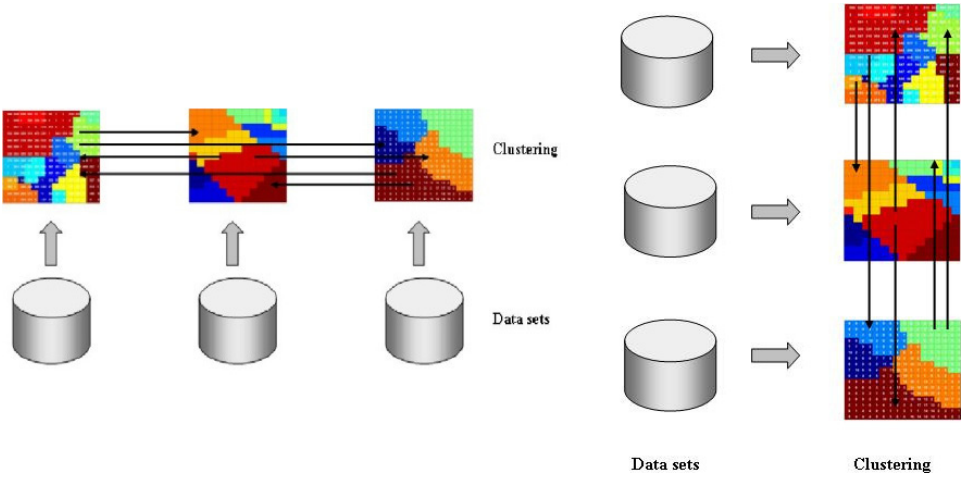


Fig. 1. A general scheme of horizontal (left) and vertical (right) collaboration using SOM.

the same and is equal to N. The collaboration between two subsets is established through an interaction coefficient which describes the intensity of the interaction. In general, $\alpha_{[ii]}^{[jj]}$ and $\beta_{[ii]}^{[jj]}$ assume non-negative values. The higher the value of the interaction (collaboration) coefficients is, the stronger the collaboration between the corresponding datasets will be. In this paper, we will estimate the interaction coefficients during the collaboration phase of the algorithm. The main idea of the horizontal collaboration between different SOM is that if an observation from the iith dataset is projected on the jth neuron in the ii-map, then that same observation in the jjth dataset will be projected on the same j neuron of the jjth map or one of its neighboring neurons. In other words, *neurons that corresponds to different maps should capture the same observations.* To accommodate the collaboration mechanism in the optimization process, the objective function of the SOM is expanded into the form

$$R_H^{[ii]}(\chi, w) = \alpha_{[ii]}^{[jj]} \sum_{i=1}^{N} \sum_{j=1}^{|w|} K_{\sigma(j,\chi(x_i))}^{[ii]} \|x_i^{[ii]} - w_j^{[ii]}\|^2 + \sum_{jj=1, jj\neq ii}^{P} \beta_{[ii]}^{[jj]} \sum_{i=1}^{N} \sum_{j=1}^{|w|} (K_{\sigma(j,\chi(x_i))}^{[ii]} - K_{\sigma(j,\chi(x_i))}^{[jj]})^2 \|x_i^{[ii]} - w_j^{[ii]}\|^2, \quad (5)$$

where P represents the number of datasets (or the classifications), N the number of observations, $|w|$ is the number of prototype vectors from the ii SOM map (the number of neurons). $\chi(x_i)$ is the assignment function which allows to find the Best Matching Unit (BMU), it selects the neuron with the closest prototype from the data x_i using the Euclidean distance.

$$\chi(x_i) = \arg\min(\|x_i - w_j\|^2).$$

$\sigma(ij)$ represents the distance between two neurons i and j from the map, and it is defined as the length of the shortest path linking cells i and j on the SOM map.

$K_{\sigma(i,j)}^{[cc]}$ is the neighborhood function on the $SOM[cc]$ map between two cells i and j. The nature of the neighborhood function $K_{\sigma(i,j)}^{[cc]}$ is identical for all the maps, but its value varies from one map to another: it depends on the closest prototype to the observation that is not necessarily the same for all the SOM maps.

The value of the collaboration parameter α is determined during the first phase of the collaboration step, and $\beta = \alpha^2$. This parameter allows determining the importance of the collaboration between each two datasets, i.e., to learn the collaboration confidence between all datasets and maps. Its value belongs to [1−10], it is 1 for the neutral link, when no importance to collaboration is given, and 10 for the maximal collaboration within a map. Its value varies after each iteration during the collaboration step. In the case of the horizontal collaborative learning, as shown in Algorithm 1, the value of the collaboration confidence parameter depends on topological similarity between both collaboration maps. To compute the collaborated

Algorithm 1 The horizontal collaboration algorithm

Random the collaboration matrix $\alpha_{[ii]}^{[jj]}$

1. Local step:

for $t = 1$ to N_{iter} **do**

For each $DB[ii]$, $ii = 1$ to P:

Find the prototypes minimizing the classical SOM:

$$w^* = \arg\min_{w} \left[\sum_{i=1}^{N} \sum_{j=1}^{|w|} K_{\sigma(j,\chi(x_i))}^{[ii]} \| x_i^{[ii]} - w_j^{[ii]} \|^2 \right]$$

2. Collaboration step:

For the horizontal collaboration of the $[ii]$ map with the $[jj]$ map:

Collaboration Phase 1:

Update the prototypes of the $[ii]$ map minimizing the objective function of the horizontal collaboration using the expression (6).

Collaboration Phase 2:

The confidence exchange parameter is adapted using the following expression:

$$\alpha_{[ii]}^{[jj]}(t+1) = \alpha_{[ii]}^{[jj]}(t) + \frac{\sum_{i=1}^{N} \sum_{j=1}^{|w|} K_{\sigma(j,\chi(x_i))}^{[ii]}}{2 \sum_{i=1}^{N} \sum_{j=1}^{|w|} (K_{\sigma(j,\chi(x_i))}^{[ii]} - K_{\sigma(j,\chi(x_i))}^{[jj]})^2}$$

$$\text{with } K_{ij} = (K_{\sigma(j,\chi(x_i))}^{[ii]} - K_{\sigma(j,\chi(x_i))}^{[jj]})^2$$

$$\text{and } \beta \leftarrow \alpha^2$$

end for

prototypes matrix, we use gradient optimization technique, we obtain the following expression:

$$w^{*[ii]} = \arg\min_{w} [R_H^{[ii]}(\chi, w)] \tag{6}$$

with:

$$w_{jk}^{*[ii]}(t+1) = w_{jk}^{*[ii]}(t) + \frac{\sum_{i=1}^{N} K_{\sigma(j,\chi(x_i))}^{[ii]} x_{ik}^{[ii]} + \sum_{jj=1, jj \neq ii}^{P} \sum_{i=1}^{N} \alpha_{[ii]}^{[jj]} L_{ij} x_{ik}^{[ii]}}{\sum_{i=1}^{N} K_{\sigma(j,\chi(x_i))}^{[ii]} + \sum_{jj=1, jj \neq ii}^{P} \sum_{i=1}^{N} \alpha_{[ii]}^{[jj]} L_{ij}},$$

where

$$L_{ij} = (K_{\sigma(j,\chi(x_i))}^{[ii]} - K_{\sigma(j,\chi(x_i))}^{[jj]})^2.$$

Indeed, during the collaboration with a SOM map, the algorithm takes into account the prototypes of the map and its topology (the neighborhood function). The horizontal collaboration algorithm is presented in Algorithm 1.

3.2. *Topological vertical collaboration*

In the case of vertical collaborative clustering, contrarily to the horizontal case, we deal with different datasets where all patterns are described in the same feature space. We establish communication at the level of prototypes of the datasets that are defined in the same feature space. The basic idea of collaboration in this case is the following: a neuron j of iith SOM map and the same neuron j of the jjth map should be very similar using the Euclidean distance. In other words, *neurons that corresponds to the different maps should represent groups of similar observations*. The proposed objective function governing a search for structure in the iith dataset is

$$R_V^{[ii]}(\chi, w) = \alpha_{[ii]}^{[jj]} \sum_{i=1}^{N} \sum_{j=1}^{|w|} K_{\sigma(j,\chi(x_i))}^{[ii]} \|x_i^{[ii]} - w_j^{[ii]}\|^2 + \sum_{jj=1, jj\neq ii}^{P} \beta_{[ii]}^{[jj]} \sum_{i=1}^{N^{[ii]}} \sum_{j=1}^{|w|} (K_{\sigma(j,\chi(x_i))}^{[ii]} - K_{\sigma(j,\chi(x_i))}^{[jj]})^2 \|w_j^{[ii]} - w_j^{[jj]}\|^2, \quad (7)$$

where P represents the number of datasets, N the number of observations of the iith dataset, $|w|$ is the number of prototype vectors from the ii-SOM map and which is the same for all the maps. We will estimate the coefficients of collaboration during the collaboration phase, as same as we did in the horizontal case. Using the gradient optimization procedure, we obtain the following formulas to compute the prototypes matrix:

$$w^{*[ii]} = \arg\min_{w} [R_V^{[ii]}(\chi, w)] \quad (8)$$

with:

$$w_{jk}^{*[ii]}(t+1) = w^{*[ii]}(t) + \frac{\sum_{i=1}^{N} K_{\sigma(j,\chi(x_i))}^{[ii]} x_{ik}^{[ii]} + \sum_{jj=1, jj\neq ii}^{P} \sum_{i=1}^{N^{[ii]}} \alpha_{[ii]}^{[jj]} L_{ij} w_{ik}^{[jj]}}{\sum_{i=1}^{N} K_{\sigma(j,\chi(x_i))}^{[ii]} + \sum_{jj=1, jj\neq ii}^{P} \sum_{i=1}^{N} \alpha_{[ii]}^{[jj]} L_{ij}}, \quad (9)$$

where

$$L_{ij} = (K_{\sigma(j,\chi(x_i))}^{[ii]} - K_{\sigma(j,\chi(x_i))}^{[jj]})^2.$$

The learning algorithm in this case is presented by Algorithm 2.

4. Experimental Results

To evaluate our proposed collaborative approaches we applied our algorithms on several datasets of different size and complexity. We chose the following datasets:

Algorithm 2 The vertical collaboration algorithm

Random the collaboration matrix $\alpha_{[ii]}^{[jj]}$

1. Local step:

for $t = 1$ to N_{iter} **do**

For each $DB[ii]$, $ii = 1$ to P:

Find the prototypes minimizing the classical SOM:

$$w^* = \arg\min_k \left[\sum_{i=1}^{N} \sum_{j=1}^{|w|} K_{\sigma(j,\chi(x_i))}^{[ii]} \|x_i^{[ii]} - w_j^{[ii]}\|^2 \right]$$

2. Collaboration step:

For the vertical collaboration of the $[ii]$ map with the $[jj]$ map:

Collaboration Phase 1:

Update the prototypes of the $[ii]$ map minimizing the objective function of the horizontal collaboration using the expression (6).

Collaboration Phase 2:

The confidence exchange parameter is adapted using the following expression:

$$\alpha_{[ii]}^{[jj]}(t+1) = \alpha_{[ii]}^{[jj]}(t) + \frac{\sum_{i=1}^{N} \sum_{j=1}^{|w|} K_{\sigma(j,\chi(x_i))}^{[ii]} \|x_i^{[ii]} - w_j^{[ii]}\|^2}{2 \sum_{i=1}^{N} \sum_{j=1}^{|w|} (K_{\sigma(j,\chi(x_i))}^{[ii]} - K_{\sigma(j,\chi(x_i))}^{[jj]})^2 \|w_j^{[ii]} - w_j^{[jj]}\|^2}$$

$$\text{with } K_{ij} = (K_{\sigma(j,\chi(x_i))}^{[ii]} - K_{\sigma(j,\chi(x_i))}^{[jj]})^2$$

$$\text{and } \beta \leftarrow \alpha^2$$

end for

waveform, Wisconsin Diagnostic Breast Cancer (wdbc), Isolet, Madelon and Spambase. We will give more details on the waveform dataset to illustrate the principle of the proposed approaches, especially in the validation.

As criteria to validate our approach we used the quantization error (distortion) on many maps of different sizes and the accuracy index for each SOM. The quantization error is the most used criteria to evaluate the quality of a Kohonen's topological map. This error measures the average distance between each data vector and its winning neuron (BMU). It is calculated using the following expression:

$$qe = \frac{1}{N} \sum \|x^{(i)} - w_{x_i}\|^2, \tag{10}$$

where N represents the number of data vectors and $w_{x^{(i)}}$ is the nearest prototype to the vector x_i. The values of the quantization error depends on the size of datasets and on the sizes of builded maps, so these values can alter according to the dataset.

The purity (accuracy) of the map is equal to the average purity of all the neurons. A good map SOM should have a high degree of the purity index. The purity of a neuron is the percentage of data belonging to the majority class. Assuming that the data labels set $L = l_1, l_2, \ldots, l_{|L|}$ and the prototypes set $C = c_1, c_2, \ldots, c_{|C|}$ are known, the purity of a map is expressed by:

$$\text{purity} = \sum_{k=1}^{|C|} \frac{c_k}{N} \times \frac{\max_{i=1}^{|L|} |c_{ik}|}{|c_k|}, \tag{11}$$

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

4.1. *Datasets*

- *Waveform dataset*: This dataset 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 dataset 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 1.0–1.0.
- *Madelon*: MADELON is an artificial dataset, which was part of the NIPS 2003 feature selection challenge. This is a two-class classification problem with continuous input variables. MADELON is an artificial dataset containing data points grouped in 32 clusters placed on the vertices of a five-dimensional hypercube and randomly labeled $+1$ or -1. The five dimensions constitute five informative features. 15 linear combinations of those features were added to form a set of 20 (redundant) informative features. Based on those 20 features one must separate the examples into the 2 classes (corresponding to the ± 1 labels). The order of the features and patterns was randomized. The original dataset was splitting in three parts (learning, validation and test), but we used only 2600 observations from learning set and from validation for which the classes were known.
- *Spam Base*: The SpamBase dataset is composed from 4601 observations described by 57 variables. Every variable described an email and its category: spam or

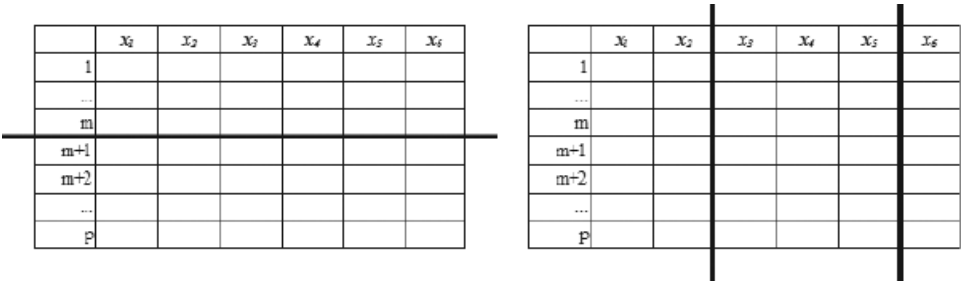


Fig. 2. Vertical (left) and horizontal (right) partitioning of data.

not-spam. Most of the attributes indicate whether a particular word or character was frequently occurring in the email. The run-length attributes (55–57) measure the length of sequences of consecutive capital letters.

4.2. *Data partitionning*

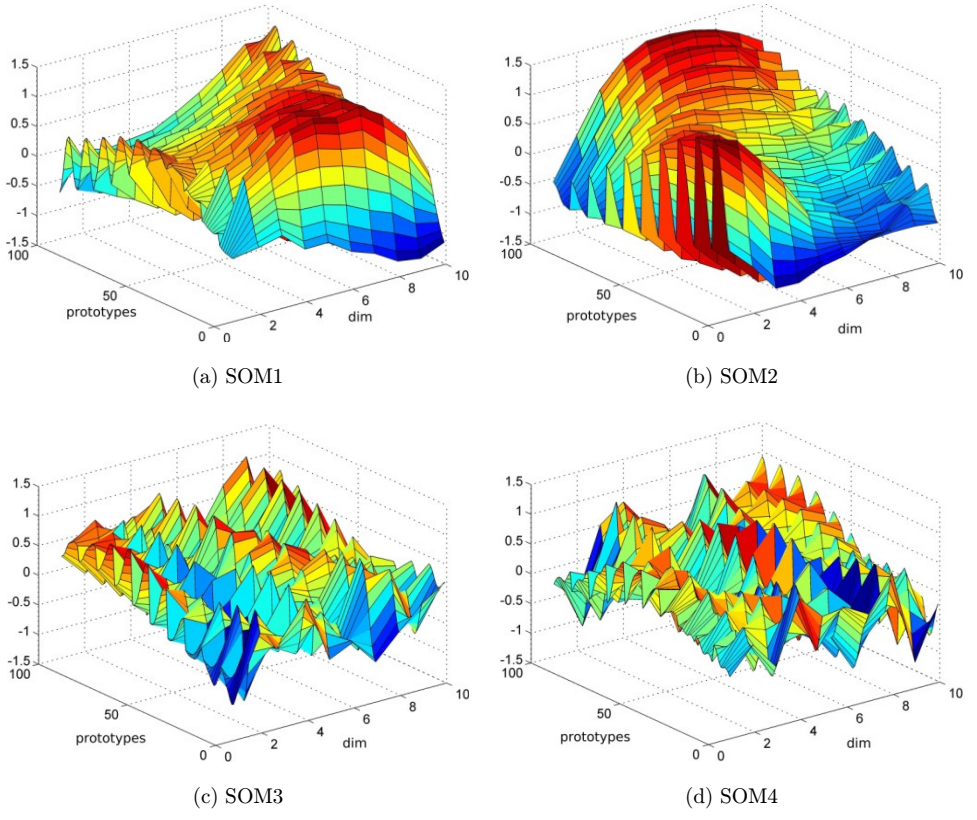
The datasets mentioned above are unified and need to be divided in subsets in order to have distributed data "scenarios". We will proceed by the vertical and horizontal partitioning (Fig. 2). In the horizontal approach we divide the datasets into subsets so that each algorithm operates on different features considering, however, the same set of individuals. In the case of vertical approach, each algorithm operates on the same features, dealing, however, with different set of individuals.

4.3. *Interpretation of the approach on the waveform dataset*

We divided the waveform dataset, of size 5000×40, into four subsets to assume a scenario of a horizontal collaboration between four sites. The first and the second part of the dataset $2 \times (5000 \times 10)$ correspond to all the relevant variables and the third and fourth part $2 \times (5000 \times 10)$ contain noisy variables. As the first and second datasets are relevant, we expect that the collaboration confidence within these datasets is bigger than the 3rd and 4th datasets.

We selected maps of size 10×10. Then we achieved the local step of the proposed approach on all four datasets which is to learn a SOM for all observations of these datasets. Figure 3 represent the prototypes vectors obtained on all the four datasets after the local step of the new learning approach. X-axis and Y-axis represent respectively the indices of variables and prototypes for these maps. Figures 3(a) and 3(b) correspond to the maps which contain the relevant variables from the waveform dataset (1-20) which are represented by the red (darker) color and have an index of purity of 81.64% and 81.5%, respectively. Knowing that the purity of the map presenting the waveform dataset before partitionning is 85.84% and the quantization error is 6.12.

We applied the second step of our algorithm to exchange the clustering information between all the maps without using the original data. Figures 4(a) and 4(b)



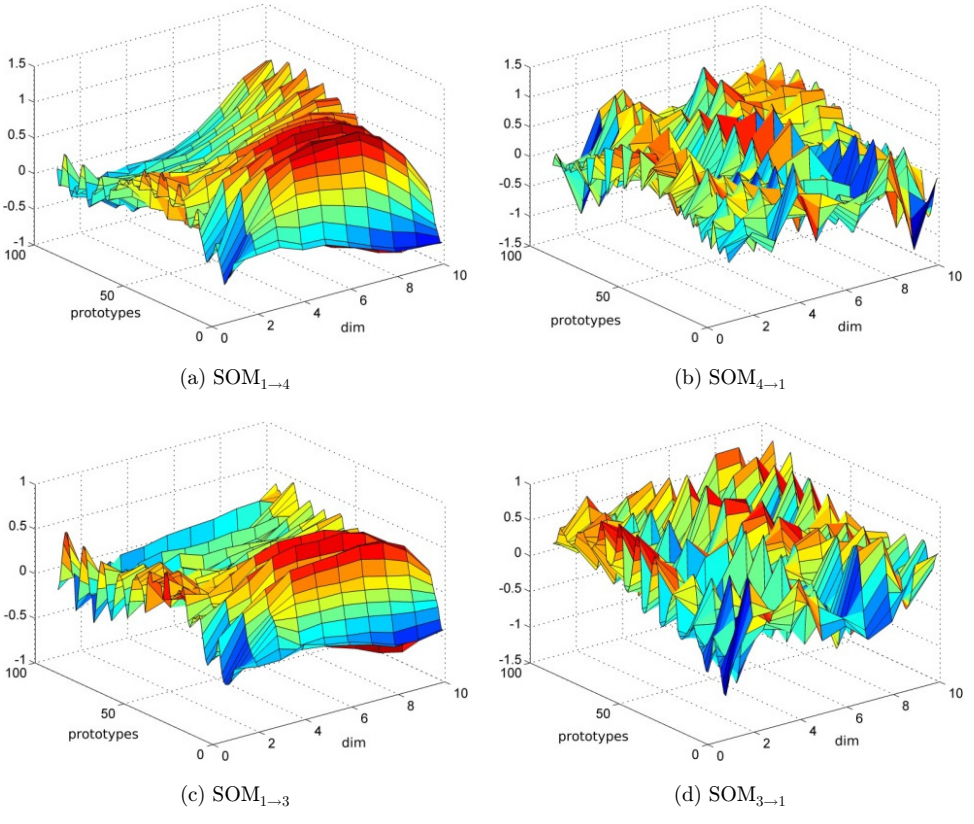
(a) SOM1 (b) SOM2

(c) SOM3 (d) SOM4

Fig. 3. Visualization of the prototypes after the first local step (classical SOM).

illustrate the collaboration between 1st and 4th datasets. After the collaboration, the purity index decreased to 78.93% because the SOM1 map (81.64%) has used the information from a noisy map (SOM4) which has very low purity index (40.21%). Contrarily, by applying the collaboration in the opposite direction, the purity index of the $SOM_{4\rightarrow 1}$ map increased to 42.45% due to the collaboration with the relevant SOM1 map (75.71% of purity). The learned collaboration confidence parameter are for the SOM_1, $\alpha = 6.03$, and for SOM_4, $\alpha = 1.34$ which means that the algorithm gives more importance to the collaboration with SOM_1 and less importance to SOM_4 map which contains noisy features.

After the collaboration of the "relevant" second dataset with the irrelevant SOM_3 map, the purity index decreased to 78.18% because the SOM_2 map (81.5%) has used the information from a noisy map (SOM_3) with a very low purity index (39.37%). Contrarily, by applying the collaboration step in the opposite direction, the purity index of the $SOM_{2\rightarrow 3}$ map increased to 41.67% due to the collaboration with the relevant SOM_2 map with a collaboration confidence parameter equals to 5.9 higher than the confidence parameter with the noisy SOM_3 map whose value is 1.2.



(a) $SOM_{1\to4}$ (b) $SOM_{4\to1}$

(c) $SOM_{1\to3}$ (d) $SOM_{3\to1}$

Fig. 4. Horizontal collaboration between the datasets 1 and 4 and between the 1st and 3rd dataset with their indices of purity.

The collaboration of a noisy map with a relevant map leads to an improvement of its quality (the purity index). The task of the horizontal collaboration is a complex problem because in an unsupervised learning process, it is difficult to identify relevant maps and we are forced to make the collaboration in both directions. Here comes the importance of learning the collaboration confidence parameters in order to give more importance to some links.

Table 1 summarizes the purities of the maps and the quantization errors before and after collaboration. As for the indices of purity, the quantization errors are improving (decreasing) after a collaboration with a more relevant map. We improve these indices due to the collaboration process and the learning of the collaboration confidence parameters. The value of each collaboration parameter is given in Table 1.

4.4. *Vertical collaboration process: Waveform dataset*

To apply vertical collaboration on waveform dataset, we divided the database into 4 subsets. The division was made randomly on the observations. We got 4 databases of

Table 1. Experimental results of collaboration approach on the waveform dataset.

Horizontal Collaboration			
Map	Purity	qe	α
SOM_1	81.64	1.98	
SOM_2	79.61	1.87	
SOM_3	47.19	2.64	
SOM_4	40.21	2.41	
$SOM_{1\to4}$	62.47	2.14	1.2
$SOM_{4\to1}$	54.63	2.27	5.9
$SOM_{2\to3}$	78.93	2.05	1.34
$SOM_{3\to2}$	41.45	2.35	6.03

size 1250×40 and we chose 1 as the value of collaboration parameter. (for the both directions). The obtained results are summarized in Table 2. We note that in most cases the purity index increases, as is the case for $SOM_{2\to1}$, $SOM_{3\to4}$, $SOM_{1\to4}$ and $SOM_{4\to1}$ and the collaboration confidence parameters are similar because all the maps are similar. As all four datasets are described in the same feature space, the purity of the maps before and after the collaboration is higher compared to the horizontal collaboration. The quantization error is also improved for the maps obtained after the collaboration with the maps having a lower quantization error.

4.5. *Validation on other datasets*

We applied the same experimental protocol on other databases and all computed indices are presented in Tables 3 and 4, for horizontal and vertical collaboration, respectively.

The size of all the used maps were fixed to 10×10. From Tables 3 and 4, we note that the purity index of the SOM maps after the horizontal collaboration increased

Table 2. Experimental results of collaboration approach on the waveform dataset.

Vertical Collaboration			
Map	Purity	qe	α
SOM_1	88.33	5.64	
SOM_2	87.75	5.83	
SOM_3	90.04	5.24	
SOM_4	88.76	5.57	
$SOM_{1\to2}$	88.06	5.62	2.2
$SOM_{2\to1}$	87.93	5.79	2.47
$SOM_{3\to4}$	90.12	5.07	2.36
$SOM_{4\to3}$	89.57	5.16	2.27
$SOM_{1\to4}$	88.46	5.59	2.41
$SOM_{4\to1}$	88.57	5.51	2.36

Table 3. Experimental results of the horizontal collaborative approach on different datasets.

Dataset	Map	Horizontal Collaboration		
		Purity	qe	α
Wdbc	SOM_1	94.95	1.99	
	SOM_2	97.27	2.07	
	$SOM_{1\to2}$	95.77	1.84	1.74
	$SOM_{2\to1}$	97.32	1.94	2.12
Isolet	SOM_1	81.20	12.61	
	SOM_2	95.12	14.45	
	$SOM_{1\to2}$	81.39	12.21	2.05
	$SOM_{2\to1}$	96.06	14.18	1.86
Madelon	SOM_1	60.88	15.58	
	SOM_2	62.64	15.50	
	$SOM_{1\to2}$	61.01	15.48	1.65
	$SOM_{2\to1}$	63.57	15.40	1.79
SpamBase	SOM_1	83.86	3.45	
	SOM_2	85.72	2.55	
	$SOM_{1\to2}$	84.17	3.23	1.92
	$SOM_{2\to1}$	83.59	2.41	1.59

for each dataset and the quantization error decreased. This is due to the use of the information from the maps related to the collaborative datasets. Also, we can note that the values of the collaboration confidence parameters are computed using the topological structure of the distant maps (distant classifications) and learning these

Table 4. Experimental results of the vertical collaborative approach on different datasets.

Dataset	Map	Vertical Collaboration		
		Purity	qe	α
Wdbc	SOM_1	96.71	90.54	
	SOM_2	97.87	67.60	
	$SOM_{1\to2}$	96.99	71.49	1.42
	$SOM_{2\to1}$	97.49	61.47	4.16
Isolet	SOM_1	98.85	8.19	
	SOM_2	98.46	8.76	
	$SOM_{1\to2}$	79.54	8.34	1.93
	$SOM_{2\to1}$	98.30	8.78	2.04
Madelon	SOM_1	69.71	61.23	
	SOM_2	69.87	61.15	
	$SOM_{1\to2}$	74.57	59.59	2.26
	$SOM_{2\to1}$	70.71	59.55	2.39
SpamBase	SOM_1	76.26	61.83	
	SOM_2	70.43	48.27	
	$SOM_{1\to2}$	72.28	45.98	1.47
	$SOM_{2\to1}$	69.78	36.74	4.25

parameters allows the system to detect the important collaboration links and directions and to avoid collaboration with irrelevant classification.

For the vertical collaboration experiments (Table 4), the size of all the maps is set to 10×10, except for the Isolet dataset whose map size is 5×5.

For the Wdbc dataset, we note that the purity index of the first SOM map after the collaboration has improved. Contrarily, the purity of the second SOM map after the collaboration decreased. We also note that the quantization error of the first and second map has improved after the collaboration. For the Isolet dataset, we do not observe any improvement on the maps obtained after the collaboration compared with that before. The purity of the maps and the quantization errors after the collaboration are improved for the Madelon dataset. For the Spam dataset, the quantization error has improved. For the vertical collaboration approach, these results show that the purity of maps and the quantization error is not always improved after collaborating the maps, and depends strongly on the relevance of the collaborative map (the quality of the collaborative classification) and on the confidence on this map (the collaboration parameter). This conclusion corresponds to the intuitive understanding of the principle and to the consequences of such cooperation.

5. Conclusion

The collaborative classification allows the interaction between the different sources of information for the purpose of revealing (detecting) the underlying structures and the regularities from the datasets. It can be treated as a process of consensus building where we search for a structure that is common to all the datasets. The impact of the collaboration matrix (the collaboration confidence values) over the overall effect of the collaboration is very important since in an unsupervised learning model there is no information about the data structure. In this study we proposed a methodology to learn the collaboration confidence parameters for both horizontal and vertical topological approaches. The proposed horizontal learning approach is adapted for collaboration between datasets that describe the same observations but with different variables, and in this case choosing the value of the collaboration confidence becomes very important as the datasets are in different feature spaces. Contrarily, 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. Both proposed approaches have been validated on multiple databases and the experimental results have shown very promising performance.

Several perspectives can be considered for this work as: Combining the horizontal and the vertical collaborative approach in order to design a new collaborative hybrid

approach; fusing all the classifications obtained after the collaboration and building a consensus classification for all the distant sites.

References

1. A. Strehl, J. Ghosh and C. Cardie, Cluster ensembles — a knowledge reuse framework for combining multiple partitions, *J. Mach. Learn. Res.* **35** (2002) 83−617.
2. J. C. da Silva and M. Klusch, Inference in distributed data clustering, *Eng. Appl. Artif. Intell.* **19**(4) (2006) 363−369.
3. W. Pedrycz, Collaborative fuzzy clustering, *Pattern Recogn. Lett.* **23**(14) (2002) 1675−1686.
4. N. Grozavu, M. Ghassany and Y. Bennani, Learning confidence exchange in collaborative clustering, *The 2011 Int. Joint Conf. Neural Networks (IJCNN)*, 31 July 2011−5 August 2011, pp. 872−879.
5. N. Grozavu and Y. Bennani, Topological collaborative clustering, in *Australian Journal of Intelligent Information Processing Systems* (*AJIIPS*), **12**(3) (2010), Machine Learning Applications (Part I).
6. W. Pedrycz, Fuzzy clustering with a knowledge-based guidance, *Pattern Recogn. Lett.* **25** (4) (2004) 469−480.
7. W. Pedrycz and P. Rai, Collaborative clustering with the use of fuzzy C-means and its quantification, *Fuzzy Set. Syst.* **159**(18) (2008) 2399−2427.
8. B. Depaire, R. Falcón, K. Vanhoof and G. Wets, PSO driven collaborative clustering: A clustering algorithm for ubiquitous, *Environ. Sci.* **15**(1) (2008) 49−68.
9. T. Kohonen, *Self-organizing Maps* (Springer-Verlag Berlin, 1995).
10. R. Falcon, G. Jeon, R. Bello and J. Jeong, Learning collaboration links in a collaborative fuzzy clustering environment, in *Proc. Artificial Intelligence 6th Mexican Int. Conf. Advances in Artificial Intelligence, MICAI'07* (Springer-Verlag, Berlin, Heidelberg, 2007), pp. 483−495.
11. C. M. Bishop, M. Svensen and C. K. I. Williams, GTM: The generative topographic mappings, *Neural Comput.* **10**(1) (1998) 215−234.
12. J. J. Verbeek, N. Vlassis and B. J. A. Krose, Self-organizing mixture models, *Neurocomputing* **63** (2005) 99−123.
13. F. Zehraoui and Y. Bennani, New self-organizing maps for multivariate sequences procing, *International Journal of Computational Intelligence and Applications* **5**(4) (2005) 439−456.