2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Diversity Analysis in Collaborative Clustering

Nistor Grozavu Guénaël Cabanes Younès Bennani

Abstract— **The aim of collaborative clustering is to reveal the common structure of data which are distributed on different sites. The topological collaborative clustering, based on Self-Organizing Maps (SOM) is an unsupervised learning method which is able to use the output of other SOMs from other sites during the learning. This paper investigates the impact of the diversity between collaborators on the collaboration's quality and presents a study of different diversity indexes for collaborative clustering. Based on experiments on artificial and real datasets, we demonstrated that the quality and the diversity of the collaboration can have an important impact on the quality of the collaboration and that not all diversity indexes are relevant for this task.**

I. Introduction

Collaborative Clustering is an emerging problem in data mining for the analysis of distributed datasets. Indeed, the current rise of real-time communication network such as the Internet and the recent advances in distributed resources in networks lead to new classes of problems. We focus here on the problem of multi-site datasets, in particular when the information stocked in some or all sites is confidential to the others. For example, we could be interested in the cluster analysis of a collection of datasets at different sites (banks, stores, medical organizations, administrations) describing the same individuals with different information, i.e. with different descriptors or variables. However, because of the confidentiality of each dataset, it is impossible to use all the data in one analysis. A local clustering is nevertheless possible in each site without breaking the confidentiality rules.

To solve this issue, previous studies [1], [2], [3] proposed the use of a Collaborative Clustering. They have shown that it is possible to use the output of a local analysis to improve the result of the clustering on another site, in order to produce an accurate view of the global hidden structure in different datasets without having direct access to the data. The aim of the Collaborative Clustering is to distribute the clustering process and merge the different results without sharing the data among different centers. In these methods, during the collaboration step, we do not need the distant datasets, we only need the result of the clustering of this dataset. Thus, each site uses its own dataset and the clustering information from distant datasets. The final partition is then as close as possible to the one we would have obtained with a centralized dataset. Pedrycz et al. [1], [2] proposed a collaborative method based on the K-Means algorithm, whereas the work of Grozavu et al. [3] is based on the learning of a modified Self-Organizing Maps (SOM) [4] to produce a Topological Collaborative Clustering.

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

The Collaborative Clustering straightly depends on some parameters which can have an important impact on the final results [3]. This is the case of the collaboration confidence matrix, which weight the influence of the collaborator on the final clustering [5]. This confidence matrix is critical in the case of collaboration, because setting in advance the strength of the collaboration for each collaboration link can degrade the final results if not set correctly. In an unsupervised collaborative learning case, no knowledge is available and usually this parameter is just set to one to avoid an unconformity to the collaborative dataset. However, estimating the optimal values could greatly improve the quality of the collaboration. In this paper, we investigate the impact of the diversity between collaborators on the quality of the collaboration and the potential role of this measure to find optimum parameters for the confidence matrix. The goal is to obtain insights on the benefit of diversity measures for a selective collaboration.

The notion of diversity starts to be increasingly used for different tasks in machine learning. In Ensemble Learning, diversity measures can be used to evaluate and improve the accuracy of a classifier [6], [7], [8], [9] or a clusterer [10], [11], [12], [13] ensemble. The main idea of an ensemble of classifiers is that each member of the ensemble is not perfect and can make errors [11]. However, different classifiers make different errors and it is possible to complement each classifier with the others, which makes errors on different objects. A global consensus between the classifiers is then reached to obtain the final partition, for example using a majority vote. The diversity of the classifier outputs is therefore a vital requirement for the success of the ensemble. Intuitively, we want the ensemble members to be as correct as possible, and in case they make errors, these errors should be on different objects [11].

We think that the same idea must be applicable in Collaborative Clustering. We therefore investigate here the link between the diversity between two potential collaborators and the accuracy gained during the collaboration process. As there is no consensus on the best diversity index to use, we also tested and compared seven different diversity measures.

The rest of this paper is organized as follows: in Section 2 we present the Topological Collaborative Clustering algorithm used in our study. Section 3 introduces the diversity indexes. In section 4 we present the experimental results. Finally the paper ends with a conclusion and future works in section 5.

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

II. Topological Collaborative Clustering

According to the structure of datasets to collaborate, there are three main types of collaboration learning principle: horizontal, vertical and hybrid collaboration. The vertical collaboration is to collaborate the clustering results obtained from different datasets described by the same variables, but having different objects. In the case of horizontal clustering, all datasets are described by the same observations but in different feature spaces: the same number of objects but a different number of variables. The hybrid collaboration is not more than a combination of the both horizontal and vertical collaboration.

In this work, we are specifically interested in horizontal collaborations. Horizontal collaboration is the most difficult one, since in such cases, the groups of data are described in different spaces: each dataset is described by different variables, but has the same objects (samples) as other datasets. In this case the problem is *how to collaborate the clusters derived out of a set of classifications from different characteristics?* and *how to manipulate the collaborative/confidence parameter where no information is available about the distant clustering?*

In the Topological Collaborative Clustering, each dataset is clustered with a Self-Organizing Map (SOM). To simplify the formalism, the maps built from various datasets will have the same dimensions (number of neurons) and the same structure (topology). The main idea of the horizontal collaboration principle between different SOM is that if an observation from the ii-th dataset is projected on the j-th neuron in the ii - map, then that same observation in the jj-th dataset will be projected on the same j-th neuron of the jj-th map or one of its neighboring neurons. In other words, *neurons that correspond to different maps should capture the same observations*. Therefore, an additional term reflecting the principle of collaboration is added to the classical SOM objective function. This function is adapted/weighted by a collaborative parameter in order to represent the confidence and the cooperation between the $[ii]$ classification and $[jj]$ classification. A new collaboration step is also added to estimate the importance of the collaboration, during the collaborative learning process. To compute the relevance of the collaboration, two parameters are introduced: the first one is to adapt the distant clustering information, and the second is for weighting the collaborative clustering link (the map which receive information about the distant map).

Formally, the following new objective function is composed of two terms:

$$R_H^{[ii]}(W) = R_{Quantiz}(W) + R_{Collab}(W)$$

with

$$R_{Quantiz}(W) = \sum_{jj=1,jj\neq ii}^{P} \alpha_{[ii]}^{[jj]} \sum_{i=1}^{N}\sum_{j=1}^{|w|} \mathcal{K}_{\sigma(j,\chi(x_i))}^{[ii]} \|x_i^{[ii]} - w_j^{[ii]}\|^2$$

and

$$R_{Collab}(W) = \sum_{jj=1,jj\neq ii}^{P} \beta_{[ii]}^{[jj]} \sum_{i=1}^{N}\sum_{j=1}^{|w|} \left(\mathcal{K}_{\sigma(j,\chi(x_i))}^{[ii]} - \mathcal{K}_{\sigma(j,\chi(x_i))}^{[jj]}\right)^2 * \|x_i^{[ii]} - w_j^{[ii]}\|^2$$

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.

$$\chi(x_i) = argmin\left(\|x_i - w_j\|^2\right)$$

$\sigma(i,j)$ 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.
$\mathcal{K}_{\sigma(i,j)}^{[cc]}$ is the neighborhood function on the $SOM[cc]$ map between two cells i and j.

The nature of the neighborhood function $\mathcal{K}_{\sigma(i,j)}^{[cc]}$ 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 to determine the importance of the collaboration between each two datasets, i.e. to learn the collaboration confidence between all datasets and maps [5]. 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 each iteration during the collaboration step.

The value of the collaboration confidence parameter depends on topological similarity between the both collaboration maps. In this case, one cannot use the prototypes vectors to compute this parameter because of the different feature spaces.

To compute the collaborated prototypes matrix, a gradient optimization is used as follow:

$$w^{*[ii]} = \arg\min_{w}\left[R_H^{[ii]}(\chi, w)\right] \tag{1}$$

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:

$$L_{ij} = \left(K^{[ii]}_{\sigma(j,\chi(x_i))} - K^{[jj]}_{\sigma(j,\chi(x_i))} \right)^2$$

Indeed, during the collaboration with a SOM map, the algorithm takes into account the prototypes of the map and its topology (the neighborhood function).

III. Diversity Measures

In Ensemble Learning, because of the relationship between the diversity of the ensemble and the ensemble performance, diversity measures is therefore helpful in designing the individual classifiers, the ensemble, and choosing the combination method.

Several diversity indexes have been proposed for this tasks, both for classification [6], [7], [8], [9] and clustering [10], [12], [13], [11] ensembles, as well as different way of using theses diversity index to improve the consensus function. The general result is that the diversity of the ensemble is indeed related to the accuracy of the ensemble. A diversity not too low neither too high is preferable. However, the definition of the diversity index is still difficult and the effect of the diversity remains difficult to quantify [6]

In this paper, we address the question of the use of the diversity for a different task. In unsupervised collaborative methods we don't try to find a consensus between several partitions, but the aims is to find the best collaboration between several clustering during the learning.

We define the diversity between two potential collaborators as the difference between the two partitions obtained separately from each of these collaborators on their own dataset. A low diversity means that the two datasets (representing the same objects in two different spaces) are partitioned in a same way by the two clustering algorithms. A high diversity means that the two dataset are partitioned in a very different way, either because of differences in the two clustering methods used or because of intrinsic difference in the data representation in the two different spaces. In our study, any high diversity were due to a difference in the data space, because we used the same algorithm to partition both datasets. To compute the diversity index we used several well-known indexes of similarity between two data partitions. These indexes are usually based on the agreement between the two partitions, i.e. each pair of object should be either in the same cluster in both partitions or in different clusters in both partitions.

In the following, we note $P1$ and $P2$ the two partitions we wish to compare. We define a_{11} as the number of object pairs belonging to the same cluster in $P1$ and $P2$, a_{10} denotes the number of pairs that belong to the same cluster in $P1$ but not in $P2$, and a_{01} denotes the pairs in the same cluster in $P2$ but not in $P1$. Finally, a_{00} denotes the number of object pairs in different clusters in $P1$ and $P2$.

A. Rand index

The Rand index [14] is one of the most used index. It can be defined as follow:

$$Rand = \frac{a_{00} + a_{11}}{a_{00} + a_{01} + a_{10} + a_{11}} \quad (2)$$

However, this index does not take into account the fact that the agreement between partitions could arise by chance alone. This could greatly bias the results for higher values of concordance [15].

B. Adjusted Rand index

To solve this issue, [15] proposed the Adjusted Rand index, which gives the overall concordance of two partitions taking into account that the agreement between them could appear by chance.

$$AdjustedRand = \frac{a_{00} + a_{11} - n_c}{a_{00} + a_{01} + a_{10} + a_{11} - n_c} \quad (3)$$

where:

$$n_c = \frac{N(N^2+1) - (N+1)\sum n_i^2 - (N+1)\sum n_j^2 + \sum\sum \frac{n_{ij}^2}{N}}{2(N-1)} \quad (4)$$

with N the total number of objects, n_i the number of objects belonging to the cluster i of $P1$, n_j the number of objects belonging to the cluster j of $P2$ and n_{ij} the number of object in cluster i in $P1$ and j in $P2$.

Here n_c is the agreement we would expect to arise by chance alone.

C. Jaccard index

The Jaccard index [16] follows the same idea as the Rand index, without taking into account the number of object pairs in different clusters in both $P1$ and $P2$.

$$Jaccard = \frac{a_{11}}{a_{01} + a_{10} + a_{11}} \quad (5)$$

D. Wallace's coefficient

Wallace's coefficient [17] can be more informative than Adjusted Rand by providing a directional information about the partition relation. It can be defined as:

$$W_{P1\to P2} = \frac{a_{11}}{a_{11} + a_{10}} \text{ and } W_{P2\to P1} = \frac{a_{11}}{a_{11} + a_{01}} \quad (6)$$

Note that $W_{P1\to P2} \neq W_{P2\to P1}$.

E. Adjusted Wallace index

For the same reason as the Adjusted Rand index, [18] proposed the used of the expected Wallace index under independence (Wi) to make sure the agreement is not due to chance alone.

$$Wi_{P1\to P2} = \frac{1}{N(N-1)} \sum_{i}^{|P2|} n_i(n_i - 1) \quad (7)$$

with N the total number of objects and n_i the number of objects in cluster i of the partition $P2$.

The Adjusted Wallace index is then defined as:

$$AW_{P1\to P2} = \frac{W_{P1\to P2} - Wi_{P1\to P2}}{1 - Wi_{P1\to P2}} \quad (8)$$

F. Normalized Mutual Information

Other type of indexes can also be used, for example based on the information theory.

The Normalized Mutual Information index, for example, can be used to compute the shared information between two partitions [19]:

$$NMI = \frac{-2\sum_{ij} n_{ij} log\frac{n_{ij}N}{n_i n_j}}{\sum_i n_i log\frac{n_i}{N} + \sum_j n_j log\frac{n_j}{N}} \quad (9)$$

with N the total number of objects, n_i the number of objects belonging to the cluster i of $P1$, n_j the number of objects belonging to the cluster j of $P2$ and n_{ij} the number of object in cluster i in $P1$ and j in $P2$.

G. Variation of Information

The Variation of Information is another index based on the information theory. This coefficient establishes how much information is included in each partitions, and how much information one partition gives about the other [20].

$$VI = -2\sum_{ij} \frac{n_{ij}}{N} log\frac{n_{ij}N}{n_i n_j} - \sum_i \frac{n_i}{N} log\frac{n_i}{N} - \sum_j \frac{n_j}{N} log\frac{n_j}{N} \quad (10)$$

with N the total number of objects, n_i the number of objects belonging to the cluster i of $P1$, n_j the number of objects belonging to the cluster j of $P2$ and n_{ij} the number of object in cluster i in $P1$ and j in $P2$.

IV. Experimental results

To evaluate the impact of the diversity on the collaborative clustering we used four datasets of different size and complexity (see Section IV-A).

A. Datasets

We performed several experiments on four datasets (one artificial and three real) from the UCI Repository of machine learning datasets [21].

- *Waveform dataset* - This artificial dataset consists of 5000 instances divided into 3 classes. The original base included 40 variables, 19 of which are 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 dataset has 569 instances with 32 variables (ID, diagnosis, 30 real-valued input variables). Each data observation is labelled 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 to 1.0.

- *Spam Base* - The SpamBase dataset is composed from 4601 instances 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.

B. Estimation of quality

As a criteria to estimate the quality of the collaboration, we computed the gain in accuracy with collaboration, in comparison to without.

A common way to assess the usefulness of a clustering result is indirect validation, whereby clusters are applied to the solution of a problem and the correctness is evaluated against objective external knowledge. This procedure is defined by [22] as "validating clustering by extrinsic classification", and has been followed in many other studies. Thus, to adopt this approach we need labelled data sets, where the external (extrinsic) knowledge is the class information provided by labels. Hence, if the clustering methods find significant clusters in the data, these will be reflected by the distribution of classes. Thus a purity score can be expressed as the percentage of elements of the assigned class in a cluster.

The accuracy of a SOM is equal to the average purity of all the neurons. A good SOM should have a high degree of the accuracy index. The accuracy of a neuron 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 accuracy of a map is the following:

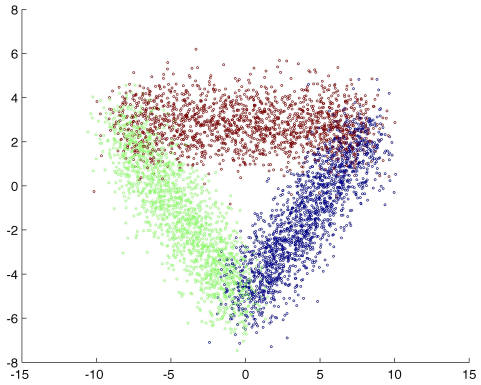
$$accuracy = \sum_{k=1}^{|C|} \frac{c_k}{N} \times \frac{max_{i=1}^{|L|}|c_{ik}|}{|c_k|} \quad (11)$$

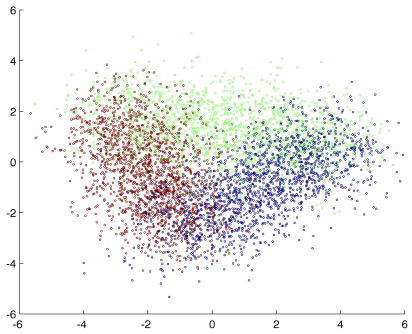
where $|c_k|$ is the total number of data associated with the neuron c_k, and $|c_{ik}|$ is the number of observations of class l_i which are associated to the neuron c_k and N - the total number of observations (data).

C. Experimental results on the Waveform dataset

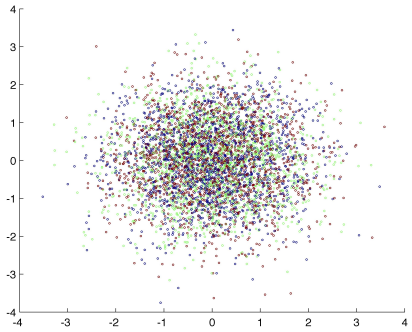
The waveform dataset is structured in 3 classes (Figure 1) and the first 20 variables from the waveform dataset correspond to relevant features and the last twenty variables - to noisy variables. From this figure it is easy to see that the data distribution in subset db1 is much better (easy to identify

Fig. 1. Waveform dataset distribution





(a) dataset db1



(b) dataset db6

Fig. 2. Data distribution of 4 datasets to collaborate

classes) compared to the subsets db6 containing only noisy variables. Intuitively, even if the diversity between a relevant dataset and a noisy dataset is high, the collaboration should not increase the results. The presence of noisy and acurate variables can therefore be used to generate "good" and "bad" collaborators and predict the diversity between collaborators and its effect.

To simplify the interpretation of the collaboration principle, in this example, we firstly assume a scenario of a collaboration between 10 sites. We divided the basic waveform dataset size 5000×40 in ten subsets as following: the first subset containing variables 1 to 4, the second dataset is composed of variables 5 to 8, and so on. So the first five subsets (db1, db2, db3, db4, db5) are composed of relevant waveform variables and the last five subsets (db6, db7, db8, db9, db10) are composed of noisy variables. Figure 2(a) represents the data visualization of a relevant subsets (db1) and Figure 2(b) represents a noisy subsets (db6). The visualization was obtained using a Principal Component Analysis (PCA) on these subsets and the colors represent the real class data distribution.

One of the challenges in the unsupervised collaborative clustering is the choice of the collaborator. As described in section 3 a diversity measure should be computed before the collaboration as a measure of the quality of the potential collaborator. In this toy example, we know that the SOM trained on the noisy datasets are "bad" collaborators as they don't have any relevant information to share, whereas the SOM trained on relevant variables are reliable collaborators.

We first learned a SOM for all of these datasets with a 10×10 map size. Then we computed all the diversity indexes introduced in section 3 on six pairs of subsets (see Table 1). These results (Table 1) represent the diversity measures computed for classifications obtained from the relevant datasets (db2/db3 and db3/db4), diversity computed between relevant vs noisy classifications (db2/db8 and db4/db9), and finally the diversity obtained for classifications issued from noisy waveform subsets only (db7/db8 and db9/db10).

The Rand, Jaccard, Wallace's, Adjusted Wallace and Variation of Information measures doesn't give a good indication about the diversity between the classifications because they are very similar for all the comparisons. For example, for the Rand index it is hard to point out that the diversity between the db2 and db3 (both relevant classifications) is much higher compared to the diversity between db4 and db9 (db4 - relevant classification and db9 - noisy classification), the difference between them is 0.12 and both greater than 0.5. More complicated is the comparison between the diversities obtained using relevant vs noisy datasets and only noisy datasets (last columns from the table 2) where the indexes are all 0.5.

Analysing the Adjusted Rand (AR) and NMI (Normal Mutual Information) measures, it is easy to note that the diversity between relevant classifications (db2/db3 and db3/db4) is much higher compared to the diversity obtained from noisy classifications (db2/db8 and db4/db9): from 0.33 to 0.2e-3. Moreover, the diversity between the noisy classifications db7/db8 and db9/db10 is also much smaller compared to the diversity obtained from relevant classifications. So, these indexes allow us easily to detect the relevant classifications to collaborate and the irrelevant classifications to not use for collaboration.

Taking into account these results, our choice in this work was to use the Adjusted Rand index. The diversity between two collaborators is represented by

TABLE I

DIVERSITY MEASURE ON THE WAVEFORM SUBSETS

Subset	Relevant datasets		Relevant vs Noisy datasets		Noisy datasets	
Diversity index	db2/db3	db3/db4	db2/db8	db4/db9	db7/db8	db9/db10
Rand	0.6707	0.7042	0.5539	0.555	0.543	0.5553
Adjusted Rand	0.2625	0.3356	0.00008	0.0002	0.00002	0.00004
Jaccard	0.3429	0.3869	0.2017	0.2008	0.2	0.2003
Wallace's coefficient	0.5079	0.5578	0.3332	0.3342	0.33	0.3334
Adjusted Wallace	0.5135	0.5581	0.3383	0.3347	0.35	0.3411
Normal Mutual Information	0.262	0.3072	0.0002	0.0006	0.0003	0.0004
Variation of Information	2.334	2.1918	3.1577	3.1631	3.168	3.1664

1 − *Adjusted Rand index*.

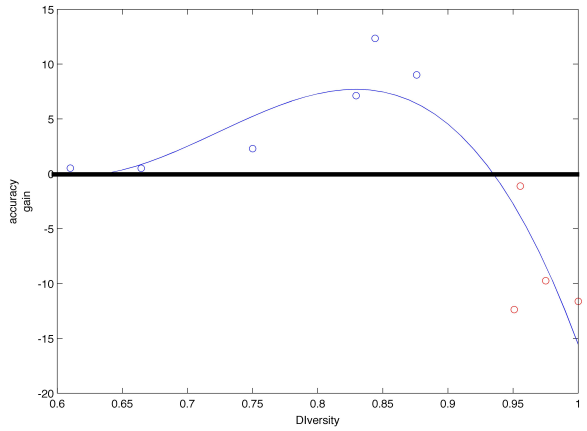
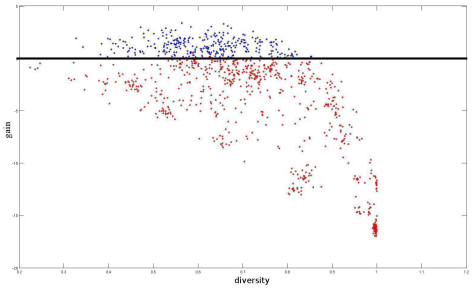


Fig. 3. The plot of diversity and the accuracy difference after collaboration

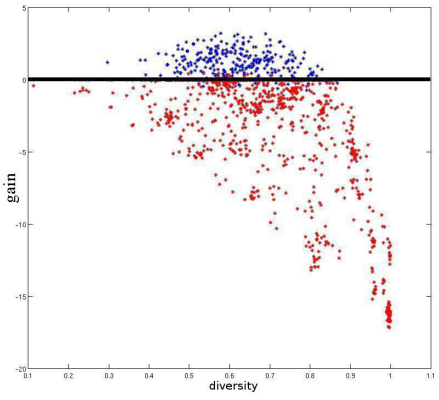
The Figure 3 presents the plot of diversity compared to the accuracy gain obtained after the collaboration on these 10 waveform subsets. The abscissa represents the Diversity index and the ordinate represents the gain obtained after the collaboration (difference between the initial accuracy and after collaboration accuracy - from -13.02% to 12.7%). Note that the real class label of the waveform dataset were used only for the validation and not for the learning of the map. The results (Figure 3) show that if the diversity is very high (diversity close to 1) the accuracy index will decrease after the collaboration, and if the diversity is small, the accuracy will not change significantly (close to 0). The accuracy index will increase more in the case of an average diversity (from 0.6 to 0.9) if the collaboration is made with a relevant map, and it can decrease in the case of the collaboration with a noisy map.

After this study, our conclusions is that if the diversity is very high (close to 0.9) the collaboration algorithm should not take into account the corresponding partition, but if the diversity is small, usually the collaboration will not produce a higher accuracy, but the partition can be used however because the accuracy index will not decrease.

In the following, for all datasets we will attempt also 1000 experiences, where each experience represents a collaboration between a fixed subset and a randomly selected subset.



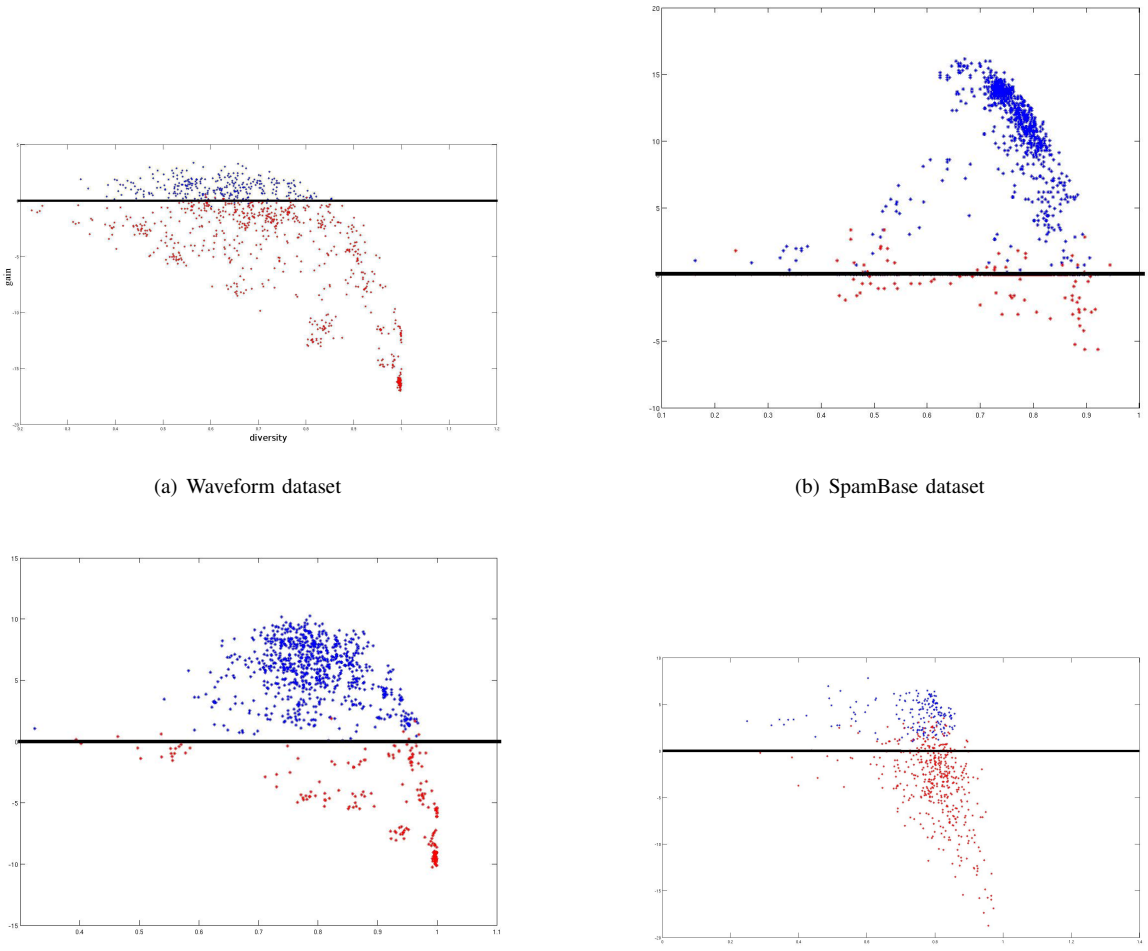
(a) waveform subset 1



(b) waveform subset 2

Fig. 4. Waveform datasets: Collaboration results between a fixed subset and 1000 randomly subsets (axe X represents the Diversity and axe Y - the Accuracy gain)

The Figure 4 represents the visualization of the obtained accuracy gain before the collaboration and the diversity between each pair of datasets for 1000 experiments randomly selected from the initial waveform dataset. Each image (Figure 4(a) and 4(b)) corresponds to the collaboration results



(a) Waveform dataset

(b) SpamBase dataset

(c) Isolet dataset

(d) Wdbc dataset

Fig. 5. Collaboration results between a fixed subset and 1000 randomly subsets (axe X represents the Diversity and axe Y - the Accuracy gain)

obtained using two fixed waveform subsets and 1000 randomly subsets for each one. The blue colour represents that the collaboration was made with a more relevant dataset and the red colour means that the collaboration was conducted with a less relevant dataset (containing noisy variables). The axis represent the diversity (axe X: from 0 to 1) and the corresponding difference between the accuracy index before the collaboration and after the collaboration (axe Y).
As it can be noted, when the diversity is small, the accuracy gain is also small, but when the diversity is high the accuracy will decrease. In this case, the high diversity means that the dataset is noisy because the fixed subset contains relevant variables, so the collaboration results depends also on the quality of the collaboration subset.

D. Experimental results on different datasets

In this section we present the results obtained on the waveform, Isolet, wdbc and spambase dataset. For all these datasets, we attempt experiments between a fixed subset and 1000 randomly selected subsets; the diversity and the accuracy gain were computed for each experiment presented in the Figure 5.

As it can be noted from the Figures 5(b), 5(c), 5(d), the results are close to those obtained on the waveform dataset (Figure 5(a)), i.e. in the case of average diversity - the accuracy gain will increase after the collaboration.

The diversity to choose depends also on the dataset. For the waveform dataset (Figure 5(a)) a good diversity is between 0.4 and 0.8, but it can be noted that the the accuracy after the collaboration of subsets having the diversity in this range can also decrease (red points in the image), that means that the collaboration were conducted between a relevant collaborator and a non-relevant collaborator (containing noisy variables). For the SpamBase dataset, the accuracy index after the collaboration will increase in the majority cases, and it will decrease when the diversity is very high, close to 0.9 as it can be noted from the Figure 5(b). The experiences attempted on the Isolet dataset represented in the Figure 5(c) shows that the accuracy will decrease when the diversity between the collaborators is in the range [0.95 - 1]. And, finally in the

case of Wisconson breast cancer dataset (Figure 5(d)), the accuracy index will decrease for a diversity situated in the range [0.85 - 1]. All these results show a similar behaviour of the collaboration results against the diversity, i.e.. **an average diversity between the collaborators allows to obtain a higher performance after the collaboration.**

It should be noted, that for all these experiences the accuracy can decrease in the case of an average diversity if the collaboration is attempted with a less relevant subset (collaborator). So, as we mentioned earlier, *the quality of the collaborator is a very important index in the case of collaborative learning* compared to consensus learning where only the diversity can be enough to conclude.

V. Conclusions

This paper focuses on Collaborative Clustering and investigates the impact of the diversity between collaborators on the collaboration's quality. We showed that only some usual diversity indexes are relevant for this task. Experiments on artificial and real datasets demonstrate the importance of the diversity on the collaboration quality. Overall, the variability of the collaboration's quality increase with the diversity. Indeed, a high diversity means that the potential collaborator achieve a very different clustering, which can be either a very good solution (in that case it is worth collaborating with it) or a very bad one (in which case the quality of the collaboration will be lower than no collaboration at all). A low diversity means that the two clustering are very similar, and none of the two collaborators will benefit on the collaborations, as they don't have any new information to share. We also shown that the quality of the clustering algorithm on its own dataset is very important for the collaboration's quality improvement regarding the diversity index. If the clustering algorithm is efficient on its dataset, it will benefit from the collaboration with collaborators not too diverse (as very diverse collaborators could perform a random clustering) nor too similar (as a similar clustering does not have any new information to share). However if the algorithm is not adapted to the dataset or if the dataset is very noisy and does not contain exploitable information, it is more probable to increase the quality of the clustering by collaborating with very diverse collaborators, to increase the change to obtain valuable information. Being able to evaluate the quality of the couple algorithm/dataset is therefore very important for Collaborative Clustering tasks.

In the future, we plan to incorporate the diversity measures as a guide for a Selective Collaborative Clustering. We wish to propose a Collaborative method between several datasets with a diversity-based weight on each of the potential collaborators, in order to optimize the final quality of the collaborative clustering.

VI. Acknowledgments

This work was supported by ANR-MN COCLICO Project.

References

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

[2] ——, "Fuzzy clustering with a knowledge-based guidance," *Pattern Recogn. Lett.*, vol. 25, no. 4, pp. 469–480, 2004.

[3] N. Grozavu and Y. Bennani, "Topological Collaborative Clustering," *in LNCS Springer of ICONIP'10 : 17th International Conference on Neural Information Processing*, 2010.

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

[5] N. Grozavu, M. Ghassany, and Y. Bennani, "Learning confidence exchange in collaborative clustering," in *IJCNN*, 2011, pp. 872–879.

[6] L. I. Kuncheva and C. J. Whitaker, "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy," *Mach. Learn.*, vol. 51, no. 2, pp. 181–207, May 2003.

[7] M. Aksela, "Comparison of Classifier Selection Methods for Improving Committee Performance," in *Multiple Classifier Systems*, ser. Lecture Notes in Computer Science, T. Windeatt and F. Roli, Eds. Springer Berlin Heidelberg, 2003, vol. 2709, pp. 84–93.

[8] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, "A new ensemble diversity measure applied to thinning ensembles," in *4th International Workshop on Multiple Classifier Systems*, 2003, pp. 306–316.

[9] D. Ruta, "Classier diversity in combined pattern recognition systems," Ph.D. dissertation, University of Paisley, 2003.

[10] X. Z. Fern and C. E. Brodley, "Random projection for high dimensional data clustering: A cluster ensemble approach," 2003, pp. 186–193.

[11] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[12] S. T. Hadjitodorov, L. I. Kuncheva, and L. P. Todorova, "Moderate diversity for better cluster ensembles," *Inf. Fusion*, vol. 7, no. 3, pp. 264–275, Sep. 2006.

[13] F. Gullo, A. Tagarelli, and S. Greco, "Diversity-Based Weighting Schemes for Clustering Ensembles," in *SDM*, 2009, pp. 437–448.

[14] W. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association.*, pp. 846–850, 1971.

[15] L. Hubert and P. Arabie, "Comparing Partitions," *Journal of the Classification*, vol. 2, pp. 193–218, 1985.

[16] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[17] D. L. Wallace, "A Method for Comparing Two Hierarchical Clusterings: Comment," *Journal of the American Statistical Association*, vol. 78, no. 383, pp. pp. 569–576, 1983. [Online]. Available: http://www.jstor.org/stable/2288118

[18] F. Pinto, J. Carrico, M. Ramirez, and J. Almeida, "Ranked Adjusted Rand: integrating distance and partition information in a measure of clustering agreement," *BMC Bioinformatics*, vol. 8, no. 1, p. 44, 2007. [Online]. Available: http://www.biomedcentral.com/1471-2105/8/44

[19] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.

[20] M. Meila, "Comparing clusterings - an information based distance," *Journal of Multivariate Analysis*, vol. 98, pp. 873–895, 2007.

[21] A. Asuncion and D. Newman, "UCI Machine Learning Repository," 2007.

[22] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.