2015 IEEE Symposium Series on Computational Intelligence

Collaborative Clustering: How to select the optimal Collaborators?

Parisa Rastin, Guénaël Cabanes, Nistor Grozavu and Younès Bennani
LIPN-UMR CNRS 7030, Université de Paris 13,
Sorbonne Paris Cité, FRANCE
Email: parisa.rastin@lipn.univ-paris13.fr

***Abstract*—The aim of collaborative clustering is to reveal the common underlying structure of data spread across multiple data sites by applying clustering techniques. The idea of Collaborative Clustering is that each collaborator share some information about the segmentation (structure) of its local data and improve its own clustering with the information provided by the other collaborators. This paper analyses the impact of the Quality of the potential Collaborators to the quality of the collaboration for a Topological Collaborative Clustering Algorithm based on the learning of a Self-Organizing Map. Experimental analysis on four real vector data-sets showed that the diversity between collaborators impact the quality of the collaboration. We also showed that the internal indexes of quality are a good estimator of the increase of quality due to the collaboration.**

I. Introduction

The current growth in real-time communication networks leads to new categories of problems. One of these new problems is the distribution of information between multiple locations and owners. For example, there are data distributed across different sites (banks, supermarkets, medical organizations, administrations) describing the same people with different information (i.e. with different variables). The analysis of distributed data-sets requires appropriate methods, particularly where the different sites can not share data directly for privacy reasons. Collaborative Clustering addresses this issue by using the clusters from remote data (the collaborators) to improve the clustering of local data [1], [2]. If all the data cannot be used in a single analysis, a local clustering is nevertheless possible in each site without breaching, for example, confidentiality rules. The idea of the Collaborative Clustering is that every collaborators shares information on the clustering (i.e. the structure) of its local data and improves its own clustering with information provided by other collaborators. As the actual data are not shared, confidentiality is preserved.

While most distributed data clustering [3], [4] produce a consensus taking into account all their data-sets, the fundamental concept of the collaborative clustering is that each algorithms operate locally on each data-set, then collaborate by exchanging information about the local data structure [5], [6], [7], [8]. Collaborative clustering is divided into two phases: a local phase and a phase of collaboration. The local phase would apply a clustering algorithm based on prototypes, locally and independently on each database. The phase of collaboration aims to collaborate each of the databases with 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 confidentiality, i.e. to collaborate databases themselves. At the end of the two phases, all the local clustering will be enriched.

Mainly, there are three different types of collaborative learning: horizontal, vertical and hybrid Grozavu2011. The vertical collaboration is to collaborate the clustering results obtained from different data sets described by the same variables with different objects. In horizontal clustering we deal with the same patterns and different feature spaces. The hybrid collaboration is a combination of the both horizontal and vertical collaboration. In this work we focus on Horizontal Topological Clustering. Grozavu *et al.* [9] proposed a Topological Collaborative Clustering method based on the learning of a Self-Organizing Map (SOM) [10]. They showed that the potential collaborators are not equivalent for the collaboration. Indeed, if the collaborator is not adequately chosen, the resulting partition can be of lesser quality than the local clustering (without collaboration). This variability is still not clearly understood.

In this paper we address the problem of the choice of the collaborator. More precisely, as some collaborators can decrease the quality of the collaboration, we investigated several methods to characterize the collaborators in order to be able to predict the potential quality of a collaborator for the collaboration. We tested the impact of several indexes of Diversity and internal Quality of the collaborators to the gain of quality after collaboration. The Diversity indexes measure the similarity between the two partitions before collaboration, whereas the internal Quality indexes measure the compactness and the separability of the partition proposed by the collaborator. Diversity and Quality indexes are known to be important for Ensemble Clustering [11], [12], which is a problem related to Collaborative Clustering.

The rest of the paper is organized as follows: the Horizontal Topological Collaborative Clustering algorithm used in this paper is described and discussed in Section II. In Section III we define the notion of Diversity and we present the different indexes used in the experiments, then the experimental protocol and the obtained results are described. Section IV focus on the Internal Quality of the collaborators, we present several indexes and the experiments performed to test their impact on the final quality of the collaboration before describing and analyzing the obtained results. Finally, the paper ends with a conclusion and some future works.

978-1-4799-7560-0/15 $31.00 © 2015 IEEE
DOI 10.1109/SSCI.2015.117

II. Topological Collaborative Clustering

According to the structure of data-sets 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 data-sets described by the same variables, but having different objects. In the case of horizontal clustering, all data-sets 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 data-set is described by different variables, but has the same objects (samples) as other data-sets. 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 data-set is clustered with a Self-Organizing Map (SOM). To simplify the formalism, the maps built from various data-sets 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 data-set is projected on the j-th neuron in the ii - map, then that same observation in the jj-th data-set 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:

$$R_H^{[ii]} = R_Q^{[ii]} + R_C^{[ii]}$$

with

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

and

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

where P represents the number of data-sets (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.

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

$\sigma(i,j)$ 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. This function depends on the distance between two neurons and is defined as follows:

$$K_{\sigma(i,j)}^{[cc]} = exp\left(-\frac{\sigma^2(i,j)}{T^2} \right)$$

where T is the temperature which allows to control the neighborhood influence of a cell on the map, it decreases with the T parameter. The value of T can vary between two values T_{max} and T_{min}.

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 data-sets, i.e. to learn the collaboration confidence between all data-sets and maps [2]. 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:

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

with:

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

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).

The horizontal collaboration algorithm is presented in Algorithm 1.

Algorithm 1 The horizontal collaboration algorithm

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

1. Local step:

for each map $[ii]$, $ii = 1$ to P **do**

Find the prototypes minimizing the classical SOM objective function:

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

end for

2. Collaboration step:

for each map $[ii]$, $ii = 1$ to P **do**

Update the prototypes of the $[ii]$ map minimizing the objective function $R_H^{[ii]}$ of the horizontal collaboration:

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

where:

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

end for

end for

In [6], the authors show that not all of the potential collaborators are suitable to collaborate (Table I). This table presents the purity [13] of the clustering before and after collaboration. The purity is the average proportion of the majority label in each cluster; "true" labels of the data must be known in order to compute the purity. A local clustering is noted M_x and a collaboration between x and y is noted $M_{x\to y}$ if x uses information from y and $M_{y\to x}$ if y uses information from x. It is quite clear that $M_{x\to y}$ is beneficial for x when y have a higher purity than x. However, if the collaborator have a lower purity, the purity of the collaboration usually decreases. This shows that is is important to choose carefully the best collaborators among the potential candidates. The purity is computed based on already known labels of the data (it is an External Quality Index) and cannot usually be computed as the labels are rarely known in real case clustering problem. We therefore need to find another criteria to estimate the quality of the collaborator for the collaboration.

TABLE I. EXPERIMENTAL RESULTS OF THE HORIZONTAL COLLABORATIVE APPROACH ON DIFFERENT DATA SETS [14]

data-set	Map	Purity	DB Index
Waveform	M_1	86.54	1.14
	M_2	39.5	3.75
	$M_{1\to2}$	73.24	1.73
	$M_{2\to1}$	58.34	0.91
Isolet	M_1	80.79	1.04
	M_2	93.27	0.89
	$M_{1\to2}$	81.46	0.97
	$M_{2\to1}$	92.87	0.83
Wdbc	M_1	94.02	0.87
	M_2	96.49	0.92
	$M_{1\to2}$	95.23	0.84
	$M_{2\to1}$	96.57	0.9
SpamBase	M_1	80.57	1.12
	M_2	84.95	0.95
	$M_{1\to2}$	82.84	1.06
	$M_{2\to1}$	83.79	0.92

The objective of this work is to test several criteria in order to choose the most relevant collaborators. The purpose is to find the best collaborator to collaborate with and improve the local results. For this, we tested two types of Internal indexes: Diversity between collaborators (Section III) and the individual Quality of each collaborators (Section IV).

III. IMPACT OF THE DIVERSITY BETWEEN COLLABORATORS

The Diversity is the difference between two cluster partitioning (local and collaborator). In ensemble methods, there is a relation between ensemble efficiency and ensemble diversity and a diversity measure can be useful to choose the combination method [11], [12].

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 [11], [15], [16], [17] and clustering [18], [19], [12], [20] 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 [11]

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. In order to test if the diversity between collaborators is a relevant index to choose the best collaborator, we first tried several diversity indexes to select the most discriminant index between informative and

non-informative collaborators. We also examined the different way of using these diversity index to find and improve the results. In general, a low diversity means that the two data sets (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 data sets 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 this study, any high diversity was due to a difference in the data space, because we used the same algorithm to partition both data sets.

A. Relevance of the Diversity indexes

To compute the Diversity indexes 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. Diversity among a pair of partitions can be defined as a measure that quantifies the degree of disagreement between them. A simple diversity measure consists in calculating the complement of a similarity measure S between two partitions $P1$ and $P2$: $D(P1, P2) = 1 - S(P1, P2)$.

We tested six Diversity indexes based on Similarity measures (Table II). 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 $P2$. N the total number of objects, n_i the number of objects in cluster i in $P1$, n_j the number of objects in cluster j in $P2$ and n_{ij} the number of object in cluster i in $P1$ and j in $P2$. In Adjusted Rand, n_c is the agreement we would expect to arise by chance alone using Rand index.

TABLE II. SIMILARITY MEASURES

Index	Formula
Rand [21]:	$R = \frac{a_{00}+a_{11}}{a_{00}+a_{01}+a_{10}+a_{11}}$
Adjusted Rand [22]:	$AR = \frac{a_{00}+a_{11}-n_c}{a_{00}+a_{01}+a_{10}+a_{11}-n_c}$
Jaccard [23]:	$J = \frac{a_{11}}{a_{01}+a_{10}+a_{11}}$
Wallace [24]:	$W_{P1\to P2} = \frac{a_{11}}{a_{11}+a_{10}}$
Adjusted Wallace [25]:	$AW = \frac{W_{P1\to P2} - \frac{\sum_i^{\lvert P2\rvert} n_i(n_i-1)}{N(N-1)}}{1-\frac{\sum_i^{\lvert P2\rvert} n_i(n_i-1)}{N(N-1)}}$
Normalized Mutual Information [26]:	$NMI = \frac{-2\sum_{ij} n_{ij} log \frac{n_{ij}N}{n_i n_j}}{\sum_i n_i log \frac{n_i}{N} + \sum_j n_j log \frac{n_j}{N}}$
Variation of Information [27]:	$VI = -2\sum_{ij} \frac{n_{ij}}{N} log \frac{n_{ij}N}{n_i n_j} - \sum_i \frac{n_i}{N} log \frac{n_i}{N} - \sum_j \frac{n_j}{N} log \frac{n_j}{N}$

To analyze the relevance of the different indexes, we tested how discriminant each index is between informative and non-informative collaborators. In that order, we used noisy features in the waveform data-set to manipulate the quantity of relevant information shared by the collaborator. In that case, the quality of the collaboration depends directly on the percentage of noise in the collaborators data-set. We used that property to test different diversity indexes. The waveform data-set is made of 5000 observation described by 40 features, 19 of them being random noise. We constructed ten subsets with only five features each, five subsets (db1 to db5) with informative features (Relevant data-sets) and five subsets (db6 to db10) with uninformative features (Noisy data-sets). Then we computed several diversity measures between two Relevant data-sets, a Noisy and a Relevant, and two Noisy data-sets, based on the comparison between the two SOM representations. A relevant diversity index should be high (close to 1) as soon as a Noisy data-set is involved, and low (close to 0) otherwise.

Table III presents the values of the Similarity indexes (i.e. 1-Diversity) in several cases of collaboration. As we can see, the Adjusted Rand index and the NMI (Normal Mutual Information) Index are both very low (close to zero) as soon as a Noisy data-set is involved and have a much higher value when two non-noisy data-sets collaborate. Therefore, they both are good candidate for a Diversity measure. The other indexes show little difference between collaboration with and without noisy and are probably not suitable be used as a Diversity index. In the following experiments, we chose to use the Adjusted Rand index as a Diversity index.

B. Effect of the Diversity on the quality of the collaboration

Tu evaluate the impact of Diversity among collaborators on the quality of the clustering after collaboration, we used four databases of different sizes and complexities: "Waveform", "Isolet", "WDBC" and "SpamBase" [28]. "Waveform" is describer in Section III-A. "Spambase" is a data-set with 4601 observations and 58 features, describing different types of emails (spam and non-spam). "Isolet" is a real data-set with of 7797 observations and 617 features divided into two classes. Finally, "WDBC" contains 2 classes on medical data (569 observations and 32 features).

The criteria we chose to estimate the quality of the collaboration is the gain of purity after collaboration (i.e. the difference between the purity of the local clustering before and after collaboration). For each data-set we generated 1000 pairs of collaborators by generating subsets of 4 random features, then we computed the gain of purity of each collaboration, i.e. for each pair. In [6], one member of the pair (noted "local" collaborator) was chosen to have a high purity before collaboration, the other (noted "remote" collaborator) being chosen randomly, and the gain was computed only for the high purity collaborator (over 0.8). Figure 1 shows the results for each data-sets, each dot represent a collaboration. Blue points represent collaborations where the local collaborator have a lower purity than the remote collaborator, red points represent the opposite case. It is clear from this Figure that the local collaborator should receive information from a remote collaborator with a higher purity to increase its own purity through the collaboration. In most real case in clustering problems we don't have the true labels of the data and it is no possible to compute the Purity index, however we can see here that the Diversity between Collaborators can provide some information about the quality of the Collaboration. Indeed,

TABLE III. DIVERSITY MEASURE ON THE WAVEFORM SUBSETS

Subset	Relevant data-sets		Relevant vs Noisy data-sets		Noisy data-sets	
Diversity index	db2/db3	db3/db4	db2/db8	db4/db9	db7/db8	db9/db10
Rand	0.6707	0.7042	0.5539	0.555	0.5430	0.5553
Adjusted Rand	0.2625	0.3356	0.00008	0.0002	0.0000	0.0000
Jaccard	0.3429	0.3869	0.2017	0.2008	0.2000	0.2003
Wallace's coefficient	0.5079	0.5578	0.3332	0.3342	0.3300	0.3334
Adjusted Wallace	0.5135	0.5581	0.3383	0.3347	0.3500	0.3411
Normal Mutual Information	0.2620	0.3072	0.0002	0.0006	0.0003	0.0004
Variation of Information	2.334	2.1918	3.1577	3.1631	3.1680	3.1664

when the diversity is low, the two collaborators propose very similar results and the collaboration is not informative, leading to a gain close to 0. However, as the local collaborator have a high purity, if the remote collaborator proposes a very different solution (high diversity), this solution is probably incorrect and the collaboration will decrease the purity of the local collaborator. An intermediate diversity is therefore optimal.

This is valid only if the local collaborator have a high purity. In the general case, the relationship between Diversity and gain of purity is different. We tested 1000 collaborations between collaborators chosen randomly (Figure 2). In the general case, it is clear that the Diversity is directly linked to the variability of the quality of the collaboration. A low Diversity between collaborators will lead to an useless collaboration. In the contrary, a high Diversity can potentially greatly improve the result after collaboration, when the local collaborator propose an incorrect solution and receive a very good solution from the remote collaborator, or greatly decrease the quality of the clustering after collaboration (in the opposite case), or anything between these two extremes. Actually the most important information here seems to be the quality of the collaborator. As the Purity is an external index which cannot be computed most of the time, we investigated the impact of several Internal Quality index to predict the gain in purity after collaboration.

IV. IMPACT OF THE INTERNAL QUALITY OF THE COLLABORATORS

In this section we analyse the link between the proportion of relevant information in a data-set and the Internal Quality of a clustering on these data. Then we show how Internal Quality indexes and gain of purity after collaboration are related to each other.

A. relationship between quantity of information and Internal Quality

We used the percentage of noisy features as an indicator of the proportion of relevant information in a data-set. Noisy data-set are less informative than noise-free data-set, because noisy features contain random values. Noise is not the only cause to explain the lack of relevant information in a collaborator: unrelated description of the objects between the two collaborators can be another one (in that case the two partitions are very different). However noise is the easiest to manipulate and only depend on the internal property of the data-set.

We investigated the Correlation between the percentage of noise in the data and the internal indexes, in order to find the best index to predict the percentage of relevant information (i.e. here the percentage of noise). In this experiment we studied six different internal indexes: Calinski-Harabasz, Davies-Bouldin, Krzanowski-Lai and Silhouette, as well as two SOM-specific indexes: Topological error and Quantization error (Table IV).

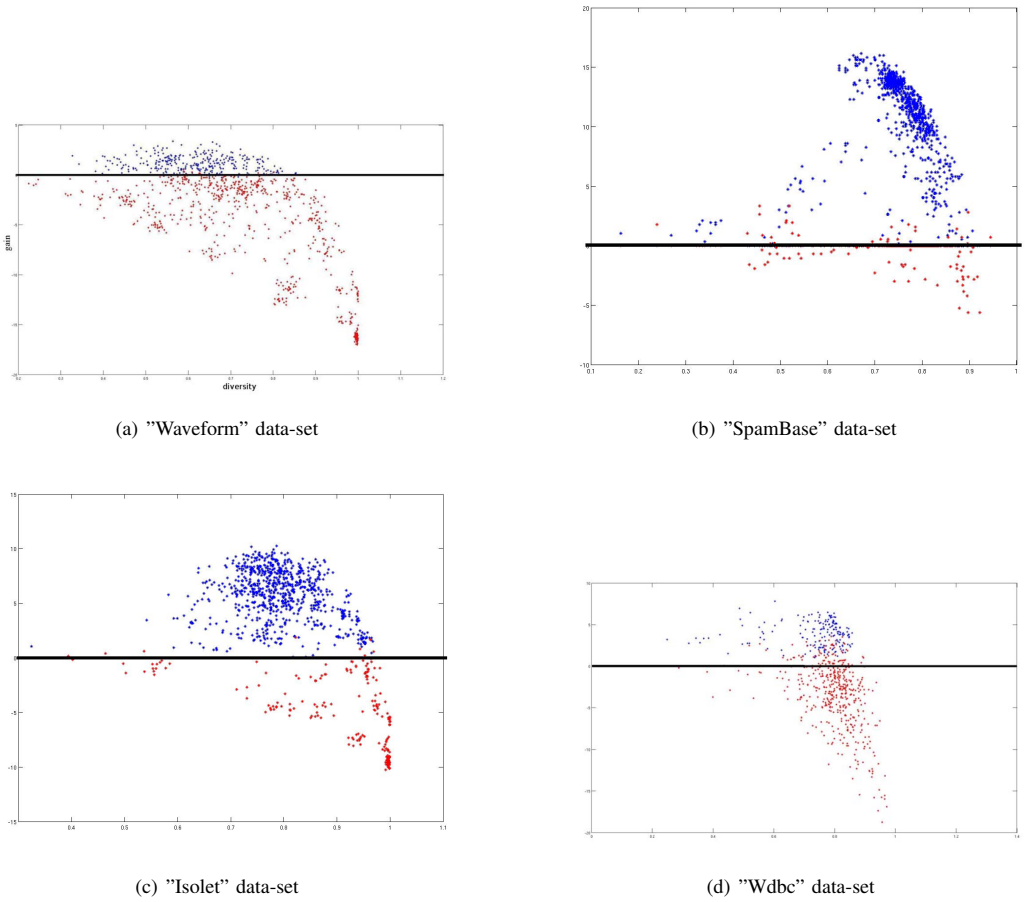
In this Table, k is the number of clusters and n is the number of data x of dimension d. W is the sum of the within-cluster variances and B is the sum of the between-cluster variances. $\overline{x_m}$ is the center of gravity of cluster m. In Silhouette, $a(i)$ is the average distance between x_i and the observations belonging to the same cluster as x_i and $b(i)$ is the lowest average distance between x_i and the observations in each other clusters.

TABLE IV. INTERNAL QUALITY INDEXES

Index	**Formula**
Calinski-Harabasz [29]:	$CH(k) = \frac{B(k)/(k-1)}{W(k))/(n-k)}$
Davies-Bouldin [30]	$DB = \frac{1}{k}\sum^{k} \max_{l \neq m} \frac{\text{avg}_{x_i \in l}\|x_i - \overline{x_l}\| + \text{avg}_i \|x_i - \overline{x_m}\|}{\|\overline{x_l} - \overline{x_m}\|}$
Krzanowski-Lai [31]:	$KL = \left\lvert \frac{W(k-1)(k-1)^{2/d} - W(k)k^{2/d}}{W(k)k^{2/d} - W(k+1)(k+1)^{2/d}} \right\rvert$
Silhouette [32]:	$S(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$
Topological error [33]:	$Te =$ proportion of data with the two closest neurons not connected
Quantization error [34]	$Qe =$ average distance between the data and their closest neurons

The experimental data-set is the waveform data-set. We generated 1000 subsets from the waveform data-set, by choosing randomly 10 features with replacement among the 40 features of waveform. As waveform contain 20 informative and 20 noisy features, the proportion of noisy features varies between 0 and 100% in the subsets. A SOM was applied of each of the subsets, and the quality indexes were computer to evaluate the SOM quality. Finally, a correlation analysis between the percentage of noise and the quality indexes were performed to find the best index to predict the noise.

Figure 3 shows the correlation between each index and the percentage of noise. There is a clear correlation with at least three indexes: Davies-Bouldinn, Calinski-Harabasz and Silhouette. Silhouette index is the best candidate to predict the proportion of relevant information due to the presence of noise in the data.



(a) "Waveform" data-set

(b) "SpamBase" data-set

(c) "Isolet" data-set

(d) "Wdbc" data-set

Fig. 1. Gain of Purity (ordinate) in function of the Diversity between collaborators (abscissa) for a collaboration of a clustering wit a high quality receiving information from a collaborator of random quality.

B. Correlation between the Internal Quality of the collaborator and the gain in Purity after collaboration

Then we conducted in 1000 collaborations between random collaborators to calculate the Pearson correlation r and the statistical significance of this correlation (t-test) between the Internal Quality Index of the remote collaborator and the gain in purity of the local collaborator after collaboration (Table V).

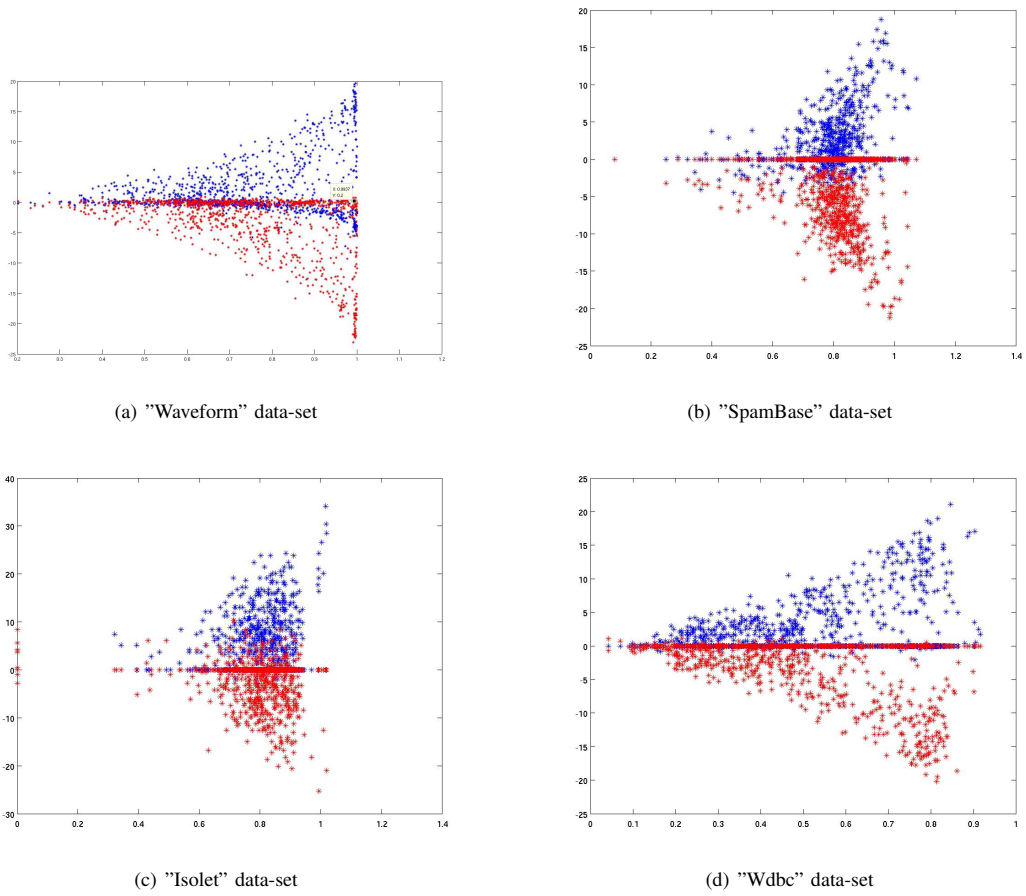
TABLE V. CORRELATION BETWEEN THE GAIN IN PURITY AND THE QUALITY OF THE COLLABORATOR FOR SEVERAL QUALITY INDEXES. r IS THE PEARSON'S CORRELATION, ALL $p < 0.001$ (T TEST).

Indices	r	p
Quantification error	-0.3395	0.0087
Topological error	-0.4497	0.0000
Silhouette	0.3915	0.0001
Davis-Bouldin	0.3936	0.0001
Calinski-Harabasz	0.3000	0.1593
Krzanowski-Lai	-0.3193	0.0405

The gain of purity is clearly significantly correlated to most internal Quality index. The best correlations are with Silhouette, Davies-Bouldin and the Topological error. Topological error is the best predictor of the gain in Purity, but this index is specific to SOM-based methods. It seems that Silhouette index is at the same time a good predictor of the proportion of noise in the data and the potential gain in Purity from a collaboration.

V. CONCLUSIONS

In this paper, we studied the impact of the collaborators Diversity and Quality on the collaboration. The results show that the Diversity between collaborators can be an important information for predicting the gain in purity due to the collaboration. However the Diversity must be completed with an estimation of the Internal Quality of the collaborators. If the Quality of the local clustering is low, any collaborator with higher Quality will improve the quality of the clustering. However, if the Quality of the local clustering is high, the optimal collaborators should have both a high Quality and an intermediate Diversity (far from 0 and 1). This is due to the fact that two good clustering solutions must be different enough to add new information to the collaboration. We showed that the Adjusted Rand index and the Normalized Mutual Information index are good candidates to estimate the diversity between Collaborators and that Silhouette index is a good estimator of



(a) "Waveform" data-set

(b) "SpamBase" data-set

(c) "Isolet" data-set

(d) "Wdbc" data-set

Fig. 2. Gain of Purity (ordinate) in function of the Diversity between collaborators (abscissa) for randomly chosen pairs of collaborators.

the quantity of noise in the data and the final Quality of the collaboration.

We plan, to pursue this work, to develop an index that combines Quality and Diversity. The idea will be to associate a confidence score to each potential collaborator and to propose new algorithms capable of multiple collaborations weighted according to these scores.

ACKNOWLEDGMENT

This work is partially funded by the ANR (French National Research Agency): project CoCLiCo ANR N 12 MONU 0001.

REFERENCES

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

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

[3] A. Strehl and J. Ghosh, "Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions," *Journal on Machine Learning Research (JMLR)*, vol. 3, pp. 583–617, Dec. 2002.

[4] J. da Silva and M. Klusch, "Inference on distributed data clustering," in *Machine Learning and Data Mining in Pattern Recognition*, ser. Lecture Notes in Computer Science, P. Perner and A. Imiya, Eds. Springer Berlin Heidelberg, 2005, vol. 3587, pp. 610–619. [Online]. Available: http://dx.doi.org/10.1007/11510888_60

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

[6] N. Grozavu, G. Cabanes, and Y. Bennani, "Diversity analysis in collaborative clustering," in *IEEE World Congress on Computational Intelligence*, 2014.

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

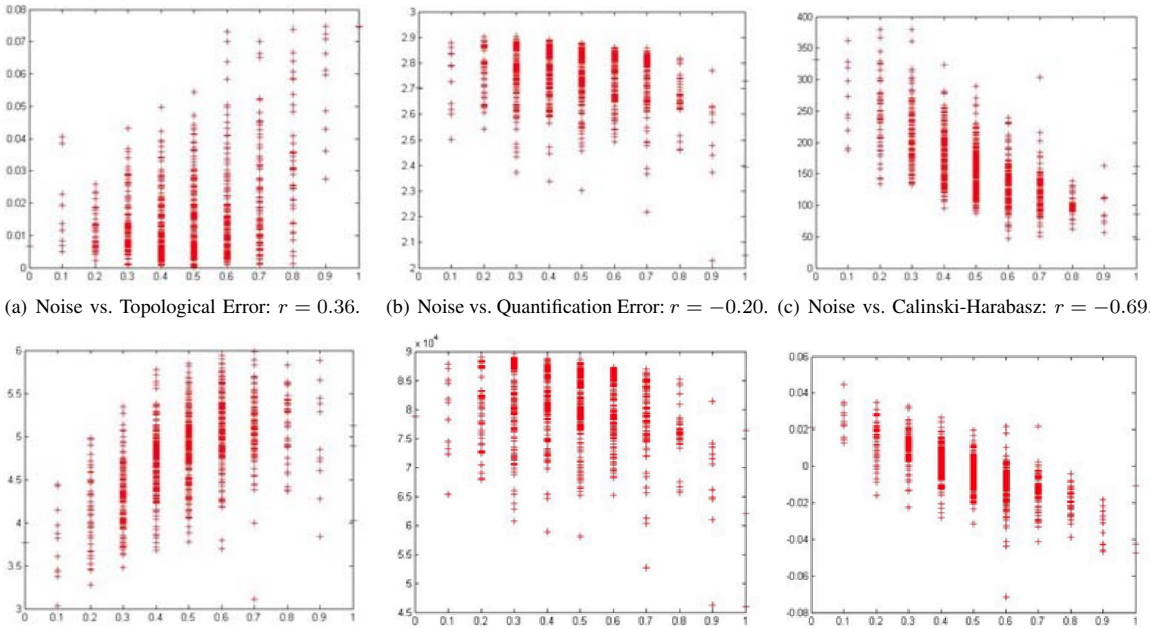
[8] M. Ghassany, N. Grozavu, and Y. Bennani, "Collaborative clustering using prototype-based techniques," *International Journal of Computational Intelligence and Applications*, vol. 11, no. 03, p. 1250017, 2012.

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

[10] T. Kohonen, *Self-Organization and Associative Memory*. Berlin: Springer-Verlag, 1984.

[11] 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.

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



(a) Noise vs. Topological Error: $r = 0.36$. (b) Noise vs. Quantification Error: $r = -0.20$. (c) Noise vs. Calinski-Harabasz: $r = -0.69$.

(d) Noise vs. Davies-Bouldin: $r = 0.56$. (e) Noise vs. Krzanowski-Lai: $r = -0.17$. (f) Noise vs. Silhouette: $r = -0.74$.

Fig. 3. Correlation between the percentage of noisy features and the quality of the clustering for several quality indexes. r is the Pearson's correlation, all $p < 0.001$ (t test).

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

[14] M. Ghassany, N. Grozavu, and Y. Bennani, "Collaborative generative topographic mapping," in *Neural Information Processing*, ser. Lecture Notes in Computer Science, T. Huang, Z. Zeng, C. Li, and C. Leung, Eds. Springer Berlin Heidelberg, 2012, vol. 7664, pp. 591–598.

[15] 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.

[16] 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.

[17] D. Ruta, "Classiï¬¬er diversity in combined pattern recognition systems," Ph.D. dissertation, University of Paisley, 2003.

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

[19] 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.

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

[21] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, Dec. 1971.

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

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

[24] 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

[25] 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

[26] 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.

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

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

[29] T. Calinski and J. Harabasz, "Dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[30] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 2, pp. 224–227, Feb. 1979.

[31] W. J. Krzanowski and Y. T. Lai, "A criterion for determining the number of groups in a data set using sum-of-squares clustering," *Biometrics*, vol. 44, no. 1, pp. pp. 23–34, 1988. [Online]. Available: http://www.jstor.org/stable/2531893

[32] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, no. 0, pp. 53 – 65, 1987. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0377042787901257

[33] K. Kiviluoto, "Topology Preservation in Self-Organizing Maps," *International Conference on Neural Networks*, pp. 294–299, 1996.

[34] T. Kohonen, *Self-Organizing Maps*. Berlin: Springer-Verlag, 2001.