New efficient clustering quality indexes

Jean-Charles Lamirel, Nicolas Dugué, Pascal Cuxac

***Abstract*— This paper deals with a major challenge in clustering that is optimal model selection. It presents new efficient clustering quality indexes relying on feature maximization, which is an alternative measure to usual distributional measures relying on entropy, Chi-square metric or vector-based measures such as Euclidean distance or correlation distance. First Experiments compare the behavior of these new indexes with usual cluster quality indexes based on Euclidean distance on different kinds of test datasets for which ground truth is available. This comparison clearly highlights altogether the superior accuracy and stability of the new method on these datasets, its efficiency from low to high dimensional range and its tolerance to noise. Further experiments are then conducted on "real life" textual data extracted from a multisource bibliographic database for which ground truth is unknown. These experiments show that the accuracy and stability of these new indexes allow to deal efficiently with diachronic analysis, when other indexes do not fit the requirements for this task.**

I. Introduction

Unsupervised classification or clustering is a data analysis technique which is increasingly widely-used in different areas of application. If the datasets to be analyzed have growing size, it is clearly unfeasible to get ground truth that permit to work on them in a supervised fashion. The main problem which then arises in clustering is to qualify the obtained results in terms of quality. A quality index is a criterion which makes possible to decide which clustering method to use, to fix an optimal number of clusters and also to evaluate or develop a new method. Many approaches have been developed for that purpose as it has been pointed out in [25] [31] [27] [1]. However, even if recent alternative approaches do exist [4] [13] [14], the usual quality indexes are mostly based on the concepts of dispersion of a cluster and dissimilarity between clusters. Computation of the latter criteria themselves relies on Euclidean distance. Most popular such indexes are the Dunn index [9], the Davis-Bouldin index [6], the Silhouette index [28], the Calinski-Harabasz index [5] and the Xie-Beni index [32]. They implement the afore mentioned concepts in slightly different ways.

Consider a dataset D made of n data points, and C, a partition in k clusters of the dataset : $C = (C_i, \cdots, C_k)$. The Dunn index (Equation 1), that has to be maximized, identifies clusters which are well separated and compacts. It combines dissimilarity between clusters and their diameters to estimate the most reliable number of clusters:

$$DU = \frac{\min_{1 \leq i < j \leq k} Diss_{DU}(C_i, C_j)}{\max_{m=1,\cdots,k} Diam(C_m)} \quad (1)$$

The dissimilarity distance usually used as the numerator is often an exhaustive dissimilarity measure between all points of distinct clusters. In general, different diameters and distances definitions can be used.

The Davies-Bouldin index (Equation 2) is similar to the Dunn index and identifies clusters which are far from each other and compacts. It should be minimized:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j=1,\cdots,k; i \neq j} \{ \frac{Diam(c_i) + Diam(c_j)}{Diss_{DB}(C_i, C_j)} \} \quad (2)$$

The dissimilarity distance used in Davies-Bouldin is a distance between centroids. It is thus far faster to compute than the Dunn Index in its usual form.

To define the Silhouette index, we first define the Silhouette width of each point (Equation 3):

$$S(i) = \frac{b(i) \cdot a(i)}{max(b(i), a(i))} \quad (3)$$

with $a(i)$ being the mean distance of point i to the other points of its cluster, and $b(i)$, the lowest average distance between i and each other cluster point set. A negative silhouette value for a given point means that the point is most suited to belong to a different cluster from the one it is allocated. Then, the Silhouette Index is computed as follows and has to be maximized:

$$SI = \frac{1}{n} \sum_{i \in n} S(i) \quad (4)$$

The Calinski-Harabasz index (Equation 5) is defined as follows:

$$CH = \frac{(n-k)}{(k-1)} \frac{BGSS}{WGSS} \quad (5)$$

with $BGSS$ the Between Group Sum of Squared that measures the dissimilarity between different clusters, and $WGSS$ the Within Group Sum of Squared that measures dissimilarity within clusters. Well separated and compact clusters should maximize this ratio.

The Xie-Beni index (Equation 6) is a compromise between the approaches provided by the Dunn index and by the Calinski-Harabasz index, and it should be minimized:

$$XB = \frac{1}{n} \frac{WGSS}{\min_{1 \leq i < j \leq k} Diss_{DU}(C_i, C_j)} \quad (6)$$

As stated in [31] [12] usual indexes have the defect to be sensitive to noisy data and outliers. In [20], we also observed that the proposed indexes are not suitable to analyze clustering results in highly multidimensional space as well as they are unable to detect degenerated clustering results. Also, these indexes are not independent of the clustering method with which they are used. As an example, a clustering method which tends to optimize WGSS, like k-means [24], will also tend to naturally produce low value for that criteria

Jean-Charles Lamirel is with the Synalp Team, LORIA, Nancy, France (email: lamirel@loria.fr).

which optimizes indexes output, but does not necessarily guarantee coherent results, as it was also demonstrated in [20]. Last but not least, as Hamerly et al. pointed out in [15], the experiments on these indexes in the literature are often performed on unrealistic test corpora made up of low dimensional data with a small number of "well-shaped" (mostly hyperspheric) embedded virtual clusters. As an example, in their reference paper, Milligan and Cooper [25] compared 30 different methods for estimating the number of clusters. They classified CH and DB in the top 10, with CH the best but their experiments only used simulated data described in a low dimensional Euclidean space. The same remark can be made about the comparison performed in [31] or in [7]. However, Kassab et al. [16] used the Reuters test collection to show that the aforementioned indexes are often unable to identify an optimal clustering model whenever the dataset is constituted by complex data which need to be represented in both high-dimensional and sparse description space, obviously with embedded non-Gaussian clusters, as is often the case with textual data. The silhouette index is considered one of the more reliable indexes among those mentioned above, especially in the case of multidimensional data, mainly because it is not a diameter-based index optimized for Gaussian context. However, like the Dunn and Xie-Beni indexes, its main defect is that it is computationally expensive, which could represent a major drawback for use with large datasets constituted by high-dimensional data.

There are also other alternatives to the usual indexes. For example, in 2009, Lago-Fernãndez et al. [18] proposed a method using *negentropy* which evaluates the gap between the cluster entropy and entropy of the normal distribution with the same covariance matrix, but again their experiments were only conducted on two-dimensional data. Other recent indexes attempts were also limited by the researchers' choice of complex parameters [31].

Our goals were manifold: to get rid of the method-index dependency problem and of the sensitivity to noise issue, and to deal with high-dimensional context while avoiding computation complexity and parameter settings. To achieve such goals, we exploited features of the data points attached to clusters instead of information carried by cluster centroids and replaced Euclidean distance with a more reliable quality estimator based on the feature maximization measure. This measure has been already successfully used by Lamirel et al. to solve complex high-dimensional classification problems with highly unbalanced and noisy data gathered in similar classes thanks to its very efficient feature selection and data resampling capabilities [22]. As a complement to this information, we shall show in the upcoming experimental section that cluster quality indexes relying on this measure do not possess any of the defects of usual approaches including computational complexity.

Section II presents a feature maximization measure and our proposed new indexes. Section III presents our first experimental context based on reference datasets. Section IV details our first results already presented in a more synthetical way in [23]. Section V presents a further step of comparative exploitation of the index on a real application of diachronic analysis of scientific litterature. Section VI draws our conclusion and ideas for future work.

II. Feature maximization for feature selection

Feature maximization is an unbiased measure which can be used to estimate the quality of a classification whether it be supervised or unsupervised. In unsupervised classification (i.e. clustering), this measure exploits the properties (i.e. the features) of cluster associated data. Its principal advantage is thus to be totally independent of the clustering method and of its operating mode.

Consider a partition C which results from a clustering method applied to a dataset D represented by a group of features F. The feature maximization measure favours clusters with a maximal feature F-measure. The feature F-measure $FF_c(f)$ of a feature f associated with a cluster c is defined as the harmonic mean of the feature recall $FR_c(f)$ and of the feature predominance $FP_c(f)$, which are themselves defined as follows:

$$FR_c(f) = \frac{\Sigma_{d \in c} W_d^f}{\Sigma_{c \in C} \Sigma_{d \in c} W_d^f} \quad FP_c(f) = \frac{\Sigma_{d \in c} W_d^f}{\Sigma_{f' \in F_c, d \in c} W_d^{f'}} \tag{7}$$

with

$$FF_c(f) = 2 \left(\frac{FR_c(f) \times FP_c(f)}{FR_c(f) + FP_c(f)} \right) \tag{8}$$

where W_d^f represents the weight of the feature f for the data d and F_c represents all the features present in the dataset associated with the cluster c. Feature Predominance measures the ability of f to *describe* cluster c. In a complementary way, Feature Recall allows to characterize f according to its ability to *discriminate* c from other clusters.

There are some important similarities between Recall and Predominance used in the proposed approach and Recall and Precision used in information retrieval. We have already exploited this analogy more thoroughly in some of our former works, like in [19], but the measures proposed here must be considered as generalizations of such information retrieval measures which are no more based on agreement but on influence of a feature materialized by a weight. Weight represents the importance of a feature for a data and furthermore for a cluster. The choice of the weighting scheme is not really constrained by the approach, but it is necessary to deal with positive values. Such scheme is supposed to figure out the significance (i.e. semantic and importance) of the feature for the data.

Feature recall is a scale independent measure but feature predominance is not. We have however shown experimentally in [22] that the F-measure which is a combination of these two measures is only weakly influenced by feature scaling. Nevertheless, to guaranty full scale independent behavior for this measure, data must be standardized.

In *supervised* context, feature maximization measure can be exploited to generate a powerful feature selection process [22]. In our *unsupervised* (clustering) context, the selection process can be used to describe or label clusters according to the most typical and representative features. This process is a non-parametrized process that uses both the capacity of F-measure to discriminate between clusters ($FR_c(f)$ index) and its ability to faithfully represent the cluster data ($FP_c(f)$ index). The set S_c of features that are characteristic of a given cluster c is belonging to a partition C is translated by:

$$S_c = \left\{ f \in F_c \mid FF_c(f) > \overline{FF}(f) \text{ and } FF_c(f) > \overline{FF}_D \right\} \quad (9)$$

where

$$\overline{FF}(f) = \Sigma_{c' \in C} \frac{FF_{c'(f)}}{|C_{/f}|} \text{ and } \overline{FF}_D = \Sigma_{f \in F} \frac{\overline{FF}(f)}{|F|} \quad (10)$$

where $C_{/f}$ represents the subset of C in which the feature f occurs.

Finally, the set of all selected features S_C is the subset of F defined by:

$$S_C = \cup_{c \in C} S_c. \quad (11)$$

In other words, the features judged relevant for a given cluster are those whose representations are better than average in this cluster, and better than the average representation of all the features in the partition, in terms of feature F-measure. Features which never respect the second condition in any cluster are discarded.

A specific concept of contrast $G_c(f)$ can be defined to calculate the performance of a retained feature f for a given cluster c. It is an indicator value which is proportional to the ratio between the F-measure $FF_c(f)$ of a feature in the cluster c and the average F-measure $\overline{FF}$ of this feature for the whole partition[1]. It can be expressed as:

$$G_c(f) = FF_c(f) / \overline{FF}(f) \quad (12)$$

The active features of a cluster are those for which the contrast is greater than 1. Moreover, the higher the contrast of a feature for one cluster, the better its performance in describing the cluster content.

Below we give an example of the operating mode of the method, on the basis of a toy-dataset encompassing two classes (*Men (M)*, *Women (F)*) described with 3 features: *Nose_Size*, *Hair_Length*, *Shoes_Size*. Figure 1 shows the source data and how the F-measure calculation of the *Shoes_Size* feature operates in the *Men* class.

As shown in Figure 2, the second step consists in calculating the average F-measure of each feature over the clusters,

[1]Using p-value highlighting the significance of a feature for a cluster by comparing its contrast to unity contrast would be a potential alternative to the proposed approach. However, this method would introduce unexpected Gaussian smoothing in the process.

Shoes_ Size	Hair_ Length	Nose_ Size	Class
9	5	5	M
9	10	5	M
9	20	6	M
5	15	5	W
6	25	6	W
5	25	5	W

FR(S,M) = 27/43 = 0.65

FP(S,M) = 27/78 = 0.35

$FF(S,M) = \frac{2(FR(S,M) \times FP(S,M))}{FR(S,M)+FP(S,M)} = 0.45$

Fig. 1. Principle of feature F-measure computation for sample data.

	F(x,M)	F(x,F)	$\overline{F(x,.)}$
Hair_ Length	0.39	0.66	0.53
Shoes_ Size	0.45	0.22	0.34
Nose_ Size	0,3	0,24	0,27

$\overline{F(.,.)}$
0.38

Fig. 2. Principle of computation of overall feature F-measure average and elimination of irrelevant features.

and the overall average F-measure for the combination of all features and all classes. In this Figure, notation $\overline{F(.,.)}$ stands here for overall average $\overline{FF}_D$ presented in (Equation 9) and notation $\overline{F(x,.)}$ stands for average of class x, which is itself computed as:

$$\overline{F(x,.)} = \Sigma_{f \in S_x} \frac{FF_x(f)}{|S_x|} \quad (13)$$

Features with F-measures that are systematically lower than the overall average are eliminated. The *Nose_Size* feature is thus removed. Remaining features (i,e. selected features) are considered active in the classes in which their F-measure is above the marginal average:

1) *Shoes_Size* is active in the *Men's* class,
2) *Hair_Length* is active in the *Women's* class.

Contrast ratio highlights the degree of activity and passivity of selected features as regards their F-measure marginal average in different classes. Figure 3 illustrates how the contrast is calculated for the example presented. In the context of this example, the contrast may be considered as a function that will virtually have the following effects:

1) Increase the length of women's hair,
2) Increase the size of the men's shoes,
3) Decrease the length of the men's hair,
4) Reduce the size of women's shoes.

As already mentioned before, the active features in a

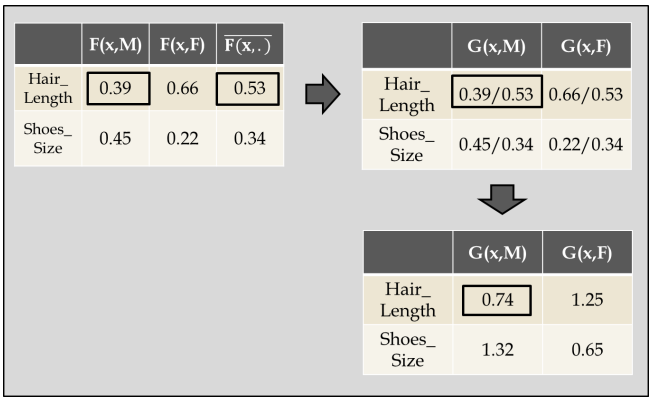


Fig. 3. Principle of computation of contrast for selected features.

cluster are selected features for which the contrast is greater than 1 in that cluster. Conversely, the passive features in a cluster are selected features present in the cluster's data for which contrast is less than unity[2]. A simple way to exploit the features obtained is to use active selected features and their associated contrast for cluster labelling as we proposed in [22]. A more sophisticated method (as we shall propose hereafter) is to exploit information related to the activity and passivity of selected features in clusters to define clustering quality indexes identifying an optimal partition. This kind of partition is expected to maximize the contrast described by eq. 12. Indeed, the more contrasted are the features, the more compact and separated the clusters. Hence, this approach leads to the definition of two different indexes.

The PC index, whose principle corresponds by analogy to that of intra-cluster inertia in the usual models, is a macro-measure based on the maximization of the average weighted contrast of active features for optimal partition. For a partition comprising k clusters, it can be expressed as:

$$PC_k = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}\sum_{f\in S_i} G_i(f) \tag{14}$$

The EC index, whose principle corresponds by analogy to that of the combination between intra-cluster inertia and inter-cluster inertia in the usual models, is based on the maximization of the average weighted compromise between the contrast of active features and the inverted contrast of passive features for optimal partition:

$$EC_k = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{\frac{|s_i|}{n_i}\sum_{f\in S_i} G_i(f) + \frac{|\overline{s_i}|}{n_i}\sum_{h\in S_i}\frac{1}{G_i(h)}}{|s_i|+|\overline{s_i}|}\right) \tag{15}$$

where n_i is the number of data associated with the cluster i, $|s_i|$ represents the number of active features in i, and $|\overline{s_i}|$, the number of passive features in the same cluster.

[2]As regards the principle of the method, this type of selected features inevitably have a contrast greater than 1 in some other cluster(s) (see eq. 9 for details).

III. Experimental data and process

To objectively evaluate the accuracy of our new indexes, we used several different datasets of varying dimensionality and size for which the optimal number of clusters (i.e. ground truth) is known in advance.

Part of the datasets came from the UCI machine learning repository [3] and is more usually exploited for classification tasks. The 4 selected UCI datasets represent mostly low to middle dimensional datasets and small datasets (except for PEN dataset which is large). The ZOO and SOY datasets which includes variables with modalities are transformed into binary files. IRIS is exploited both in standard and in binarized version to obtain clearer insight into the behavior of quality index on binary data.

VERBF is a dataset of French verbs which are described both by semantic features and by subcategorization frames. The ground truth of this dataset has been established both by linguists who studied different clustering results and by a gold standard based on the VerbNet classification, as in [29]. This binary dataset contains verbs described in a space of 231 Boolean features. It can be considered a typical middle size and middle dimensional dataset.

The R8 and R52 corpora were obtained by Cardoso Cachopo from the R10 and R90 datasets, which are derived from the Reuters 21578 collection[3]. The aim of these adjustments was to only retain data with a single label. Considering only monothematic documents and classes that still had at least one example of training and one of test, R8 is a reduction of the R10 corpus (the 10 most frequent classes) to 8 classes and R52 is a reduction of the R90 corpus (90 classes) to 52 classes. The R8 and R52 datasets, with respective size of 7674 and 9100 documents, and associated bag of words description spaces of 1187 and 2618 words, can be considered as large and high dimensional datasets.

The summary of overall datasets characteristics is provided in Table I.

We exploited 2 different usual clustering methods, namely k-means [24], a winner-take-all method, and GNG [11], a winner-take-most method with Hebbian learning. For text and/or binary datasets we also used the IGNGF neural clustering method [20] which has already been proven to outperform other clustering methods, including spectral methods [29], on this kind of data. We have reported on the method that produced the best results in the following experiments.

As class labels were provided in all datasets and considering that the clustering method could only produce approximate results as compared to reference categorization, we also used purity measures to estimate the quality of the partition generated by the method as regards to category ground truth. Following [29], we use modified purity (mPUR) to evaluate the clusterings produced, which is computed as follows:

$$mPUR = \frac{|P|}{|D|} \tag{16}$$

[3]http://www.research.att.com/∼lewis/reuters21578.html

TABLE I

DATASETS OVERALL CHARACTERISTICS (BINARIZATION OF IRIS DATASET RESULTS IN 12 BINARY FEATURES OUT OF 4 REAL-VALUED FEATURES).

	IRIS	IRIS-b	WINE	PEN	SOY	ZOO	VRBF	R8	R52
Nbr. class	3	3	3	10	16	7	12-16	8	52
Nbr data	150	150	178	10992	292	101	2183	7674	9100
Nbr feat.	4	12	13	16	84	114	231	3497	7369

where $P = \{d \in D \mid \text{prec}(c(d)) = g(d) \wedge |c(d)| > 1\}$ with D being the set of exploited data points, $c(d)$ a function that provides the cluster associated to data d and $g(d)$ a function that provides the gold class associated to data d. Clusters for which the prevalent class has only one element are considered as marginal and are thus ignored.

For the same reason, we also varied the number of clusters in a range up to 3 times that determined by the ground truth [4]. An index which gave no indication of optimum in the expected range was considered to be out-of-range or diverging index (- out-). We finally designed a process which consists in generating disturbance in the clustering results by randomly exchanging data between clusters to different fixed extents (10%, 20%, 30%) whilst maintaining the original size of the clusters. This process simulated increasingly noisy clustering results and the aim was to estimate the robustness of the proposed estimators.

IV. RESULTS

The results are presented in Tables II-IV. Some complementary information is required regarding the validation process. In the Tables, MaxP represents the number of clusters of the partition with highest mPUR value (Equation 16), or in some cases, the interval of partition sizes with highest stable mPUR value. When a quality index identified an optimal model with MaxP clusters and MaxP differed from the number of categories established by ground truth, its estimation was still considered valid. This approach took into account the fact that clustering would quite systematically produce sub-optimal results as compared to ground truth. The partitions with the highest purity values were thus studied to deal with this kind of situation. For similar reason, all estimations in the interval range between the optimal k (ground truth) and MaxP values were also considered valid. When indexes were still increasing and decreasing (depending on whether they were maximizers or minimizers) when the number of clusters was more than 3 times the number of expected classes, there were considered out-of-range (-out- symbol in Tables II-IV). The Figure 4 draws the trends of evolution of EC and PC indexes in the case of the R52 dataset. It highlights what is a suitable index behaviour (EC index) and in a parallel way what represents the out-of-range index behaviour we mentioned before (PC index).

[4]We choose too discard models of size 1 (one single cluster) in our experiments. First, the experimented indexes are not planed to produce results (or only incoherent ones) for such specific case. Second, such models are irrelevant because they correspond to trivial clustering operation.

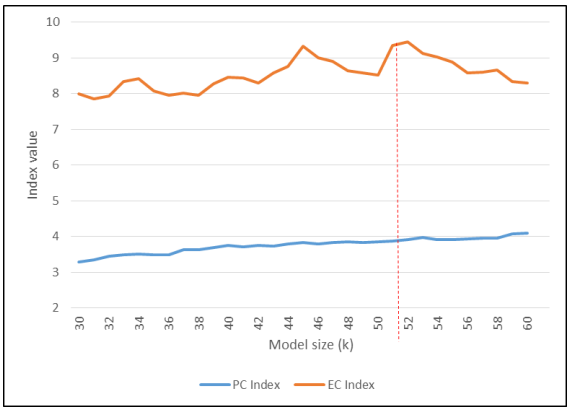


Fig. 4. Trends of PC and EC indexes on Reuters R52 dataset.

When considering the results presented in Tables II-III, it should first be noted that one of our tested indexes, the Xie-Beni (XB) index never provides any correct answers. These were either out of range (i.e. diverging) or answers (i.e. minimum value when this index was a minimizer) in the range of the variation of k, but too far from ground truth or even too far from optimal purity among the set of generated clustering models. Some indexes were in the low mid-range of correctness and provide unstable answers. This was the cases with the Davis-Bouldin (DB), Calinski-Harabasz (CH), Dunn (DU) and Silhouette (SI) indexes. With higher dimensions, these indexes were generally unable to provide any correct estimation. This phenomenon has already been observed in previous experiments with Davis-Bouldin (DB) and Calinski-Harabasz (CH) indexes [16]. Davis-Bouldin (DB) performed slightly better than average on low dimensional data. Our PC index was found to perform significantly better than average on low dimensional data but remains a better low dimensional problem estimator than a high dimensional one. Help from passive features somehow seems mandatory to estimate an optimal model in the case of high dimensional problems. Hence, the EC index which exploited both active and passive features was found to have from far the best performance, with low (Table II) or high dimensional data (Table III). According to our evaluation criteria, this index only returns wrong results in the case of the PEN dataset. However, even in this case its estimation (model of size 9) is still in the close neighbour of the optimal one (model of size 10). Additionally, the EC and PC indexes, were both found to be capable of dealing with binarized data in a transparent manner, which is not the case of some

TABLE II

OVERVIEW OF THE INDEXES ESTIMATION RESULTS ON LOW DIMENSIONAL DATA (BOLD NUMBERS REPRESENT VALID ESTIMATIONS).

	IRIS	IRIS-b	WINE	PEN	SOY	Number of correct matches
DB	2	5	**5**	7	**19**	**2**/5
CH	2	**3**	6	8	5	**1**/5
DU	1	1	8	17	8	**0**/5
SI	**4**	2	7	14	14	**1**/5
XB	2	7	-out-	19	24	**0**/5
PC	**3**	**3**	**4**	9	**16**	**4**/5
EC	**3**	**3**	**4**	9	**16**	**4**/5
MaxP	3	3	5	11	19	
Method	K-means	K-means	GNG	GNG	GNG	

TABLE III

OVERVIEW OF THE INDEXES ESTIMATION RESULTS ON AVERAGE TO HIGH DIMENSIONAL DATA (BOLD NUMBERS REPRESENT VALID ESTIMATIONS).

	ZOO	VRBF	R8	R52	Number of correct matches
DB	**8**	-out-	5	58	**1**/4
CH	4	7	**6**	-out-	**1**/4
DU	**8**	2	-out-	-out-	**1**/4
SI	4	-out-	-out-	**54**	**1**/4
XB	-out-	23	-out-	-out-	**0**/4
PC	7	18	-out-	-out-	**1**/4
EC	**7**	**15**	**6**	**52**	**4**/4
MaxP	10	12-16	6	50-55	
Method	IGNGF	IGNGF	IGNGF	IGNGF	

TABLE IV

INDEXES ESTIMATION RESULTS IN THE PRESENCE OF NOISE (UCI ZOO DATASET).

	ZOO	ZOO Noise 10%	ZOO Noise 20%	ZOO Noise 30%	Number of correct matches
DB	**8**	4	3	3	**1**/4
CH	4	5	3	3	**0**/4
DU	**8**	2	2	2	**1**/4
SI	14	-out-	-out-	-out-	**0**/4
XB	-out-	-out-	-out-	-out-	**0**/4
PC	6	4	11	**9**	**1**/4
EC	**7**	5	6	**9**	**2**/4
MaxP	10	7	10	10	
Method	IGNGF	IGNGF	IGNGF	IGNGF	

of the usal indexes namely the Xie-Beni (XI) index, and to a lesser extent, Calinski-Harabasz (CH) and Silhouette (SI) indexes. One potential explanation is that binarization process introduces some sparsity that is better dealt by our indexes than by those which use Euclidean distance.

Interestingly, on the UCI ZOO dataset, the results of noise sensitivity analysis presented in Table IV underline the fact that noise has a relatively limited effect on the operation of PC and EC indexes. The EC index was again found to have the most stable behavior in that context. The Figure 5 presents a parallel view of the different trends of EC value on non noisy and noisy clustering environment, respectively. It shows that noise tends to lower the index value in an overall way and to soften the trends relatively to changes in k value. However, the index is still able to estimate, either the optimal model in the best case, or a neighbour model in the worst case. The usual indexes do not work as well at all in the same context. For example, the Silhouette index firstly delivered the wrong optimal k values on this dataset before getting out of range when the noise reached 20% on clustering results.

The Davis-Bouldin (DB) and Dunn (DU) indexes were found to shift from a correct to a wrong estimation as soon as noise began to appear.

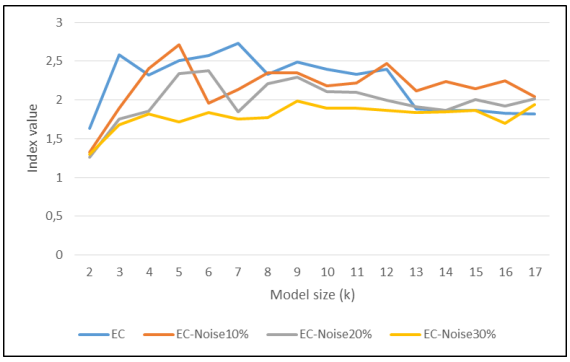


Fig. 5. Trends of EC indexes on UCI ZOO dataset with and without noise.

In all our experiments, we observed that the quality estimation depends little on the clustering method. Morever, we noted that the computation time of the index was one of the lowest among the indexes studied. As an example, for the R52 dataset, the EC index computation time was 125s as compared to 43000s for the Silhouette index using a standard laptop with 2,2 GHz quadricore processor and 8 GB of memory.

V. Comparative Exploitation of the Indexes for Diachronic Analysis

In this section, we propose a "real life" estimation of the accuracy of the cluster quality indexes, whenever they are integrated in an operationnal data mining environnement. We thus made the choice to compare their behaviour in the framework of a diachronic analysis environnement working in the context of scientific litterature. The role of such environment is to highlight different kinds of research topics changes or similarities that could occur between time periods (appearing topics, disappearing topics, splitting topics, merging topics, stable topics).

A first version of this environnement working on indexer keywords has been proposed by Lamirel [21] for demonstrating the feasabilty of a fully unsupervised approach exploiting clutering and unsupervised Bayesian reasoning between views (MVDA) for diachronic mining. A new version of this environmement working on full-text of the papers is up to now under developpement in the context of the ISTEX[5] project. A demonstrator of this new environment is already available online [6] and described in Dugué et al [8].

Figure 6 identifies the different steps of the proposed diachronic analysis process [7].

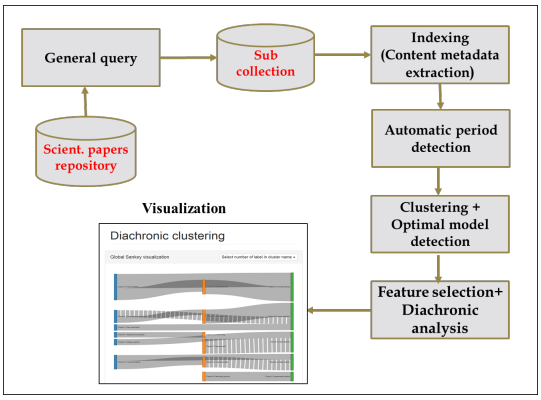


Fig. 6. General architecture of the diachronic analysis environment.

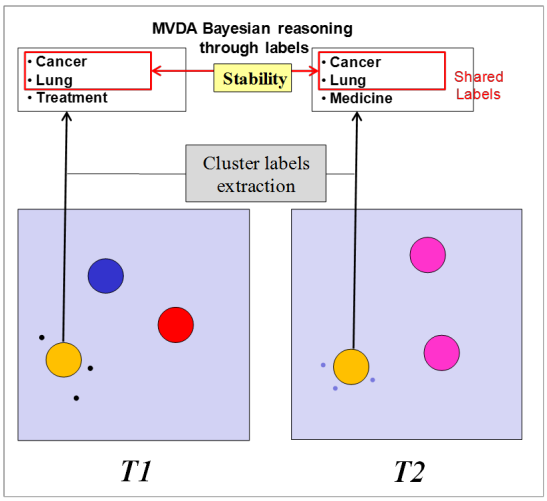


Fig. 7. Overall view of the principle of identification of topic matching between periods (the term label is used here for feature).

In this particular experiment, we first constitute a dataset of bibliographic data from the ISTEX project. The ISTEX database is queried to extract papers related to research in medical care between years 1996 and 2010. This results in a dataset of 9779 papers. A basic indexing is exploited on the full text of the documents in order to produce content metadata (single and multi-words indexing terms). The resulting index size is 8347 terms with freq. > 4. A random walk algorithm (here Walktrap [26]) exploiting the relationship between the different extracted terms and the publication years is used to extract meta-periods ([30]). At the issue of this process, three meta-periods are extracted (P1=1996-2000, P2=2000-2005, P3=2006-2010). This is between these periods that the diachronic analysis would be lead. This process should allow to monitor topic evolution in medical research across these periods. In each period, we use clustering to group documents of same topics. GNG clustering is therefore launched several times with standard Fritzke parameters settings [11] on the data of each meta-period (P1= 4054 data and 3023 terms, P2= 2012 data and 2036 terms, P3= 3713 data and 3288 terms) making varying

[5]The ISTEX project (Excellence Initiative for Scientific and Technical Information) is part of the "Investments for the Future" program initiated by the French Ministry for Higher Education and Research (MESR).

[6]See https://github.com/nicolasdugue/istex-demonstrateur

[7]Diachronic analysis and feature selection code can be found at https://github.com/nicolasdugue/istex

TABLE V

OVERVIEW OF THE INDEXES ESTIMATION RESULTS AND RESULTING QUALITY FOR DIACHRONIC ANALYSIS (THE HIGHER NUMBER OF MATCHES AND QMA, THE BETTER)

	Opt. Period P1	Opt. Period P2	Opt. Period P3	Number of temporal matches	QMA evaluation criteria
DB	-out-	-out-	-out-	0	0
CH	3	4	4	5	15.26
DU	14	20	-out-	6	8.60
SI	-out-	-out-	-out-	0	0
XB	-out-	-out-	-out-	0	0
PC	23	23	-out-	9	10.61
EC	10	6	8	**13**	**27.20**

the number of clusters in a range from 2 to 50 clusters. The different quality indexes are exploited to highlight an optimal model in each period. On this optimal model, the feature selection and contrast ([22]) method described in Section II is used to extract characteristics and salient terms for each cluster, i.e. describe cluster topics. These cluster topics descriptions are then used to compute the diachronic analysis and thus monitor the topic evolution across periods.

The diachronic analysis is made using unsupervised Bayesian reasoning [21]. Basically, signals are propagated between clusters of the different periods through the means of the salient features (Figure 7). Matching rules between clusters (topics) are based on the probability of activating a cluster of a period knowing that a cluster of an alternative period has been activated with the signal. Starting from a cluster s of a source period and issuing on a cluster t of a target period, this probability is computed as :

$$P(t|s) = \frac{\sum_{f \in S_s \cap S_t} G_t(f)}{\sum_{f \in S_t} G_t(f)} \tag{17}$$

where $G_t(f)$ is the contrast of the feature f for the cluster t (Eq. 12).

A match between a cluster of a source period and a cluster of a target period is detected if both source and target cluster generate mutual activity which is superior to the average activity they generate with the clusters of the alternative period (see [21] for more details). If such match exists, the set $S_{st} = S_s \cap S_t$ of salient features that are shared between the two aforementioned clusters is called the matching kernel.

The quality indexes have a major role in the process of diachronic analysis because the more accurate will be the choice of the optimal cluster model for each time period, the better and the more accurate will be the temporal alignment between periods and consequently the temporal matching process between those latter. Our hypothesis (that could be experimentally verified) is thus that an accurate model selection will produce the larger number of matches, with matching kernels of the largest sizes and with the highest matching probability. We consequently exploit two complementary criteria for the evaluation of the behaviour of the indexes:

- The total number of matches that can be found between the periods using the optimal clustering models provided by the index.
- A matching quality criteria (QMA) combining the number of matches, the size of the matching kernels and the matching probability. This criteria is expressed as :

$$QMA = \sum_{(i,j) \in M} |S_{ij}| * \frac{P(i|j) + P(j|i)}{2} \tag{18}$$

where M represents the set of couples of clusters of different periods for which a match has been detected.

In Table V, we give the results of our evaluation on the corpus we constitued from the ISTEX database which is related to medical care research. It highlights that, similarly to our former experiments (see Section IV), some indexes are unable to find any optimal model (being out-of-range) whatever time period is considered. Some indexes, like Dunn (DU) and our PC index, work partially, finding optimal model only in some periods. Alternatively, Calinsky-Harabasz (CH) and our EC index are able to identify optimal models in all periods.

When the evaluation criteria figuring out the quality of the diachronic analysis (i.e. temporal matches) is exploited, it turns out that the best temporal matching results are produced with the help of EC index that reached the best values both for the number of detected matches and the richness and the accuracy of the matches (QMA criteria). Additionally, despite the number of considered periods, Calinsky-Harabasz (CH), which identified optimal models for 3 periods, provided worth results than our PC index, which identified optimal models only for two periods, relatively to the number of matches. However, the exploitation of Calinsky-Harabasz (CH) index produced better quality matches than PC index as it is figured out by its higher values of QMA criteria.

VI. Conclusion

We have proposed a new set of indexes for clustering quality evaluation relying on feature maximization approach. This method exploits the information derived from features which could be associated to clusters by means of their associated data. The experiments we achieved showed that most of the usual quality estimators do not produce satisfactory results in a realistic data context and that they are additionally sensitive to noise and perform poorly with high dimensional data. Unlike the usual quality estimators, one of the main advantages of our proposed indexes is that they produce stable results in cases ranging from a low dimensional to high dimensional context and also require low computation time while easily dealing with binarized data. Their stable operating mode with clustering methods which could produce both different and imperfect results also constitutes an essential advantage.

We have performed experiments in two different contexts including a "real life" context related to diachronic analyis environment. Both experiments confirm the clear advantages of our new indexes, especially the ones of our EC quality index.

However, further experiments are required using both an extended set of clustering methods and a larger panel of high dimensional datasets to confirm this promising behavior. Especially the influence of the sparsity and the one of cluster overlapping ratio on the performance of our indexes must be more precisely evaluated. Taking a larger panel of indexes, including score functions and entropy based indexes would be also a complementary and important issue.

Additionally, we plan to test the ability of our indexes to discriminate between correct and degenerated clustering results in the context of large and heterogeneous datasets.

References

[1] Angel Latha Mary, S. and Sivagami, A.N. and Usha Rani, M.: Cluster validity measures dynamic clustering algorithms. ARPN Journal of Engineering and Applied Sciences 10(9) (2015)

[2] Arellano-Verdejo, J. and Guzmn-Arenas, A. and Godoy-Calderon, S. and Barrn Fernãndez R.: Efficiently Finding the Optimum Number of Clusters in a Dataset with a New Hybrid Cellular Evolutionary Algorithm. Computacin y Sistemas, 18(2):313–327 (2014)

[3] Bache, K. and Lichman, M.: UCI Machine learning repository (http://archive.ics.uci.edu/ml). University of California, School of Information and Computer Science, Irvine, CA, USA (2013)

[4] Bock, H.-H.: Probability model and hypothese testing in partitionning cluster analysis. In: Clustering and Classification, P. Arabie, L.J. Hubert, & G. De Soete (Eds), World Scientific, Singapore (1996), 377–453.

[5] Calinsky, T. and Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics, 3, no. 1:1–27 (1974)

[6] Davies, D. L. and Bouldin, D. W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1, no. 2:224–227 (1979)

[7] Dimitriadou, E. and Dolnicar, S. and Weingessel, A.: An examination of indexes for determining the number of clusters in binary data sets. Psychometrika, 67(1):137–159 (2002)

[8] Dugué, N. and Lamirel, J.-C. and Cuxac, P.: Diachronic'Explorer : keep track of your clusters ! Research Challenge in Information Science, 2016

[9] Dunn, J.: Well separated clusters and optimal fuzzy partitions. Journal of Cybernetics, 4:95–104 (1974)

[10] Falk, I., Lamirel J.-C., Gardent C.: Classifying French Verbs Using French and English Lexical Resources. Proccedings of ACL, Jeju Island, Korea (2012)

[11] Fritzke, B. A growing neural gas network learns topologies. Advances in Neural Information Processing Systems 7, Tesauro, G. and Touretzky, D.S. and Leen, T. K. (Ed.), 625–632 (1995)

[12] Guerra, L. and Robles, V. and Bielza, C. and Larrañaga, P.: A comparison of clustering quality indices using outliers and noise. Intelligent Data Analysis 16, 703–715 (2012)

[13] Gordon, A. D.: External validation in cluster analysis. Bulletin of the International Statistical Institute, 51(2), 353–356 (1997) Response to comments. Bulletin of the International Statistical Institute 51(3), 414–415, (1998)

[14] Halkidi, M. and Batistakis, Y. and Vazirgiannis, M.: On clustering validation techniques. Journal of Intelligent Information Systems, 17:2/3, 147–155, (2001)

[15] Hamerly, G. and Elkan, C.: Learning the K in K-Means. In Neural Information Processing Systems (2003)

[16] Kassab, R. and Lamirel, J.-C.: Feature Based Cluster Validation for High Dimensional Data. IASTED International Conference on Artificial Intelligence and Applications (AIA), 97–103, Innsbruck, Austria, February 2008.

[17] Kolesnikov, A. and Trichina, E. and Kauranne, T.: Estimating the number of clusters in a numerical data set via quantization error modeling. Pattern Recognition, 48(3): 941–952 (2015)

[18] Lago-Fernãndez, L.F. and Corbacho, F.: Using the Negentropy Increment to Determine the Number of Clusters. In Bio-Inspired Systems: Computational and Ambient Intelligence, J. Cabestany, F. Sandoval, A. Prieto, et J. M. Corchado, d. Springer Berlin Heidelberg, pp. 448–455 (2009)

[19] Lamirel, J.-C. and Francois, C. and Al Shehabi, S. and Hoffmann, M. : New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping. Scientometrics 60(3):445-462 (2004)

[20] Lamirel, J.-C. and Mall, R. and Cuxac, P. and Safi, G.: Variations to incremental growing neural gas algorithm based on label maximization. Proceedings of IJCNN 2011, 956–965, San Jose, CA, USA (2011)

[21] Lamirel, J.-C. : A new approach for automatizing the analysis of research topics dynamics: application to optoelectronics research. Scientometrics 93(1), 151–166 (2012)

[22] Lamirel, J.-C., Cuxac P., Chivukula A.S., Hajlaoui K.: Optimizing text classification through efficient feature selection based on quality metric. Journal of Intelligent Information Systems, Special issue on PAKDD-QIMIE 2013, 1–18 (2014)

[23] Lamirel, J.-C.: Reliable Clustering Quality Estimation from Low to High Dimensionnal Data. 11th Workshop on Self-Organizing Maps - WSOM 2016, Houston, TX, USA (2016)

[24] MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability (1). University of California Press, 281–297 (1967)

[25] Milligan, G.W. and Cooper, M.C.: An Examination of Procedures for Determining the Number of Clusters in a dataset. Psychometrika, 50(2):159–179 (1985).

[26] Pons, P. and Latapy, M.: Computing Communities in Large Networks Using Random Walks. Journal of Graph Algorithms and Applications, 10(2): 191–218 (2006)

[27] Rendón, E. and Abundez, I. and Arizmendi, A. and Quiroz, E.M.: Internal versus External cluster validation indexes. Internal Journal of Computers and Communications, 5(1), 27–34 (2011)

[28] Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20:53–65 (1987)

[29] Sun, L. and Korhonen, A. and Poibeau, T. and Messiant, C.: Investigating the cross-linguistic potential of VerbNet-style classification. Proceedings of ACL, 1056–1064, Beijing, China (2010)

[30] Tebbakh, A. and Dugue, N. and Lamirel, J.-C.: Feature selection and complex networks methods for an analysis of collaboration evolution in science: an application to the ISTEX digital library. 5th International Symposium ISKO-MAGHREB 2015, Hammamet, Tunisia, November 2015.

[31] Yanchi, L and Zhongmou, L. and Xiong, H. and Gao, X. and Wu, J.: Understanding of internal clustering validation measures. Proceedings of the 2010 IEEE International Conference on Data Mining, ICDM '10, 911–916.

[32] Xie, X.L. and Beni, G.: A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(8):841–847 (1991)