New quality indexes for optimal clustering model identification with high dimensional data

Jean-Charles Lamirel
LORIA
Synalp Team
Vandoeuvre les Nancy, France
Email: jean-charles.lamirel@loria.fr

Pascal Cuxac
INIST-CNRS
Department Projects and Innovations
Vandoeuvre les Nancy, France
Email: pascal.cuxac@inist.fr

***Abstract*—Feature maximization is an alternative measure to usual distributional measures relying on entropy or on Chi-square metric or vector-based measures such as Euclidean distance or correlation distance. One of the key advantages of this measure is that it is operational in an incremental mode both on clustering and on traditional classification. In the classification framework, it does not present the limitations of the aforementioned measures in the case of the processing of highly unbalanced, heterogeneous and highly multidimensional data. We shall present a new application of this measure in the clustering context for the creation of new cluster quality indexes which can be efficiently applied for a low-to-high dimensional range of data and which are tolerant to noise. We shall 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 the superior accuracy and stability of the new method.**

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 permits 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 it 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 has been pointed out in [1] [23] [24] [27]. However, even if recent alternative approaches do exist [4] [12] [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. In the next section we shall refer to the most widely used indexes which implement the afore mentioned concepts in slightly different ways.

II. Standard quality indexes

The Dunn index [8] is one of the most popular indexes. The Dunn index (DU) identifies clusters which are well separated and compact. It combines dissimilarity between clusters and their diameters to estimate the most reliable number of clusters. The Dunn index for k clusters is defined by the equation:

$$DU_k = \min_{i=1,\ldots,k} \left\{ \min_{j=i+1,\ldots,k} \left\{ \frac{diss(c_i, c_j)}{\max_{m=1,\ldots,k} diam(c_m)} \right\} \right\} \quad (1)$$

where $diss(c_i, c_j) = \min_{x \in c_i, y \in c_j} \|x - y\|$ and $diam(c_i) = \max_{x,y \in c_i} \|x - y\|$ is the intra-cluster function (or diameter) of the cluster c.

If the Dunn index is high, this means that compact and well separated clusters exist. Therefore, the maximum should be observed for k equal to the most probable number of clusters in the dataset.

The Davies-Bouldin index [6] is similar to the Dunn index and identifies clusters which are far from each other and compact. The Davies-Bouldin index (DB) is defined according to the equation:

$$DB_k = \frac{1}{k} \sum_{i=1}^{k} \max_{j=1,\ldots,k, i \neq j} \left\{ \frac{diam(c_i) + diam(c_j)}{\|c_i - c_j\|} \right\} \quad (2)$$

where, in this case, the diameter of a cluster is defined in a slightly different way as:

$$diam(c_i) = \left(\frac{1}{n_i} \sum_{x \in c_i} \|x - z_i\|^2 \right)^{1/2} \quad (3)$$

with n_i the number of data points attached to the cluster c_i and z_i the centroid of cluster c_i. Since the objective is to obtain clusters with minimum intra-cluster distances, small values for DB are interesting. Therefore, this index has to be minimized when looking for the best number of clusters.

The Silhouette index [25]: The silhouette statistic is another well-known way of estimating the number of groups in a dataset. For each point the Silhouette index (SI) computes a width depending on its membership in any cluster. 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

978-1-4673-8493-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDMW.2015.220

is allocated. The overall silhouette width is thus an average over all observations. This leads to equation:

$$SI_k = \frac{1}{k}\sum_{i=1}^{k}\frac{b_i - a_i}{\max(a_i, b_i)} \tag{4}$$

where n is the total number of points, a_i is the average distance between point i and all other points in its own cluster and b_i is the minimum of the average dissimilarities between i and points in other clusters. Finally, the partition with the highest SI is taken to be optimal.

The Calinski-Harabasz index (CH) [5]: This renowned index computes a weighted ratio between the within-group scatter and the between group scatter. Well separated and compact clusters should maximize this ratio. This leads to the equation:

$$CH_k = \frac{n-k}{k-1}\frac{BGSS}{WGSS} \tag{5}$$

with:

$$WGSS = \sum_{i=1}^{k}\sum_{x\in c_i}\|x - z_i\|^2 \tag{6}$$

and:

$$BGSS = \frac{1}{k}\sum_{i=1}^{k} n_i\|G - z_i\|^2 \tag{7}$$

where $WGSS$ represent the within-group scatter which is the sum of the squared distance between cluster centroids and their related members.

$BGSS$ stands for the between group scatter sum. Geometrically, this refers to the weighted sum of the squared distances between the centroids of the clusters and G, the barycenter of the whole set of data, the weight associated to a cluster c_k being the number n_k of elements in that cluster.

Xie-Beni index [28]: The Xie-Beni index (XI) is a compromise between the approaches provided by the Dunn index and by the Calinski-Harabasz index. It is expressed as:

$$XI_k = \frac{1}{N}\frac{WGSS}{\min_{i=1,\ldots,k}\{\min_{j=i+1,\ldots,k} diss(c_i, c_j)\}} \tag{8}$$

with N being the total number of points in the dataset.

Xie-Beni index is often exploited in fuzzy clustering. Its value must be minimized for figuring out an optimal k.

As stated in [11] [17] [27] most of the presented indexes have the defect to be sensitive to the noisy data and outliers. In [19], Lamirel et al. 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 [22], will also tend to naturally produce low value for that criteria which optimizes indexes output, but does not necessarily guarantee coherent results, as it was also demonstrated in [19]. 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" embedded virtual clusters. As an example, in their reference paper, Milligan and Cooper [23] 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 [27] or in [7]. However, Kassab et al. [16] used the Reuters test collection to shown 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 highly multidimensional 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 of highly multidimensional data.

There are also other altenatives to the usual indexes. For example, in 2009 Lago-Fernández 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. Also other recent indexes attempts were limited by the researchers'choice of complex parameters [2].

Our aim was to get rid of the method-index dependency problem and the issue of sensitivity to noise while also avoiding computation complexity, parameter settings and dealing with a highly multidimensional context. To achieve 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 highly multidimensional classification problems with highly imbalanced and noisy data gathered in similar classes thanks to its very efficient feature selection and data resampling capabilities [21]. 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 III presents a feature maximization measure and our proposed new indexes. Section IV presents our experimental context. Section V our results before section VI draws our conclusion and ideas for future work.

III. 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 data points that can be attached to their nearest cluster after analysis without prior examination of the generated cluster profiles, like centroids. 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:

$$FR_c(f) = \frac{\Sigma_{d\in c} W_d^f}{\Sigma_{c\in C}\Sigma_{d\in c} W_d^f} \quad FP_c(f) = \frac{\Sigma_{d\in c} W_d^f}{\Sigma_{f'\in F_c, d\in c} W_d^{f'}} \tag{9}$$

with

$$FF_c(f) = 2\left(\frac{FR_c(f)\times FP_c(f)}{FR_c(f)+FP_c(f)}\right) \tag{10}$$

where W_d^f represents the weight[1] 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 maximization measurement can be used to generate a powerful feature selection process [13]. In the clustering context, this kind of selection process can be defined as non-parametrized process based on cluster content in which a cluster feature is characterised using both its capacity to discriminate between clusters ($FP_c(f)$ index) and its ability to faithfully represent the cluster data ($FR_c(f)$ index). The set S_c of features which are characteristic of a given cluster c belonging to a partition C is translated by:

$$S_c = \left\{ f \in F_c \mid FF_c(f) > \overline{FF}(f) \text{ and } FF_c(f) > \overline{FF}_D \right\} \tag{11}$$

[1]The weight calculates the influence of a feature for a given data. It could be either Boolean or real-valued. An example of potential weighting scheme is given in figures 1 to 3.

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 = 0.62

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

FF(S,M) = $\frac{2(FR(S,M)\times FP(S,M))}{FR(S,M)+FP(S,M)}$ = 0.48

Fig. 1. Principle of computation of feature F-measure on example data.

where

$$\overline{FF}(f) = \Sigma_{c'\in C}\frac{FF_{c'(f)}}{|C_{/f}|} \text{ and } \overline{FF}_D = \Sigma_{f\in F}\frac{\overline{FF}(f)}{|F|} \tag{12}$$

and $C_{/f}$ 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:

$$S_C = \cup_{c\in C} S_C. \tag{13}$$

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 were 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[2]. It can be expressed as:

$$G_c(f) = FF_c(f)/\overline{FF}(f) \tag{14}$$

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 there is an example of operating mode of the method on a basis of a toy dataset including two predefined categories (i.e. classes) (*Men (M)*, *Women (F)*) described with 3 features: *Nose_Size Hair_Length*, *Shoes_Size*. Figure 1 shows the source data and also shows how the calculation of F-measure of the *Shoes_Size* feature in the *Men* class operates.

[2]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.

As shown in figure 2, the second step in the process is to calculate the marginal average of F-measure for each feature and the overall average of F-measure for the combination of all features and classes. Features with an F-measure 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 marginal average:

1) *Shoes_Size* is active in the *Men* class,
2) *Hair_Length* is active in the *Women* class[3].

The contrast ratio highlights the selected features' degree of activity/passivity with regard to their F-measure marginal average in different classes. Figure 3 illustrates how the contrast is calculated on the given example. In the context of this example, the contrast may thus be considered as a function that will virtually have the following effects:

1) Increase the length of womens' hairs,
2) Increase the size of the mens' shoes,
3) Decrease the length of the mens' hairs,
4) Reduce the size of womens' shoes.

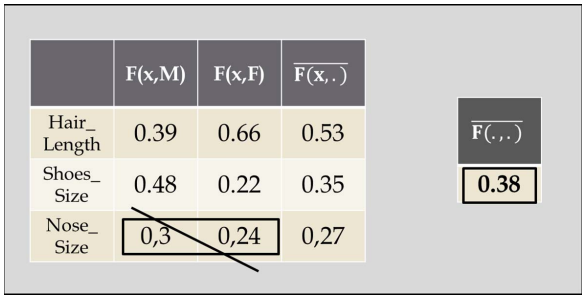

Fig. 2. Principle of computation of overall feature F-measure average and elimination of irrelevant features.

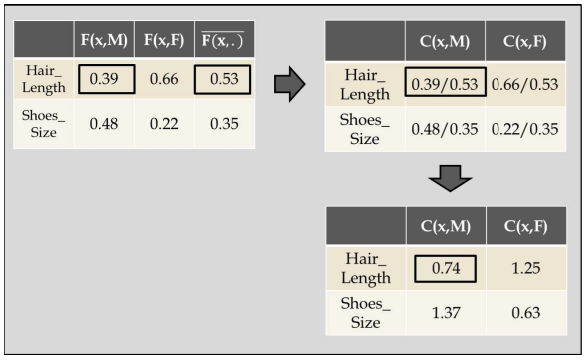

Fig. 3. Principle of computation of contrast on selected features.

[3]The method was shown in [21] to have a low sensitvity to feature scaling.

As already mentioned before, the active features in a 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 1 [4]. 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 [20]. 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. 14. 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:

$$PC_k = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}\sum_{f\in S_i} G_i(f) \tag{15}$$

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:

$$EC_k = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{\frac{|s_i|}{n_i}\sum_{f\in S_i} G_i(f) + \frac{|\overline{s_i}|}{n_i}\sum_{h\in S_i}\frac{1}{G_i(h)}}{|s_i| + |\overline{s_i}|}\right) \tag{16}$$

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.

IV. Experimental data and process

To objectively calculate 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.

A 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 dataset which includes variables with modalities was transformed into a binary file. IRIS is exploited both in standard and in binarized version to obtain clearer insight into the behavior of quality index on

[4]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. 11).

binary data.

The VERBF dataset 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 [26]. 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[5] from the R10 and R90 datasets, which are derived from the Reuters 21578 collection[6]. The aim of these adjustments was to only retain data with a single label. R8 only considers monothematic documents and classes with at least one example of training and one of testing and is a reduction of the R10 corpus (the 10 most frequent classes) to 8 classes while R52 is a reduction of the R90 corpus (90 classes) to 52 classes. R8 and R52 are large and multidimensional datasets with respective sizes of 7674 and 9100 data and an associated bag of word description spaces of 1187 and 2618 words. These datasets can be considered large and high dimensional.

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

We exploited 2 different usual clustering methods, namely k-means [22], a winner-take-all method, and GNG [10], a winner-take-most method with Hebbian learning. For text and/or binary datasets we also used the IGNGF neural clustering method [19] which has already been proven to outperform other clustering methods, including spectral methods [26], 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 [26], we use modified purity (mPUR) to evaluate the clusterings produced and this was computed as follows. Each induced cluster c was assigned the gold class (its *prevalent class*, prev(c)) to which most of its member data belonged. A data d was then said to be correctly assigned if the gold class associated it with the prevalent class of the cluster it was in. Given this, purity is the ratio between the number of correctly affected data in the clustering and the total number of data in the clustering[7]:

[5]http://web.ist.utl.pt/ acardoso/datasets/

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

[7]Clusters for which the prevalent class has only one element are considered as marginal and thus ignored.

$$mPUR = \frac{|P|}{|D|} \tag{17}$$

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.

For the same reason, we also varied the number of clusters in a range up to 3 times that determined by the ground truth. 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 obtained a process which consists of 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 aims was to estimate the robustness of the proposed estimators.

V. Results

The results are presented in tables II-III. 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 (eq. 17), 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. This kind of situation might occur when clustering identifies category subparts as separate clusters, or conversely, merges different categories into a single category. It might thus leads to an optimal number of clusters which is could be greater or lower than the expected number of categories. In our dataset sample, this could have been the case of the IRIS dataset which is known to contain 2 overlapping categories. 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-III).

When considering the results presented in table II, it should first be noted that one of our tested indexes, the Xie-Beni (XB) index, almost 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

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

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	ZOO	VRBF	R8	R52	Number of correct matches
DB	**2**	5	**5**	7	**8**	-out-	5	58	**3**/8
CH	**2**	**3**	6	8	4	7	**6**	-out-	**3**/8
DU	**2**	**2**	8	17	**8**	2	-out-	-out-	**3**/8
SI	**4**	**2**	7	14	4	-out-	-out-	**54**	**3**/8
XB	**2**	7	-out-	19	-out-	23	-out-	-out-	**1**/8
EC	**3**	**3**	**4**	**9**	**6**	18	-out-	-out-	**5**/8
PC	**3**	**2**	**4**	**9**	**7**	**15**	**6**	**52**	**8**/8
MaxP	3	3	5	11	10	12-16	6	50-55	
Method	K-means	K-means	GNG	GNG	IGNGF	IGNGF	IGNGF	IGNGF	

TABLE II
OVERVIEW OF THE INDEXES ESTIMATION RESULTS (BOLD NUMBERS REPRESENT VALID ESTIMATIONS).

	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**	**2**/4
EC	**7**	5	**6**	**9**	**3**/4
MaxP	7	7	10	10	
Method	IGNGF	IGNGF	IGNGF	IGNGF	

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

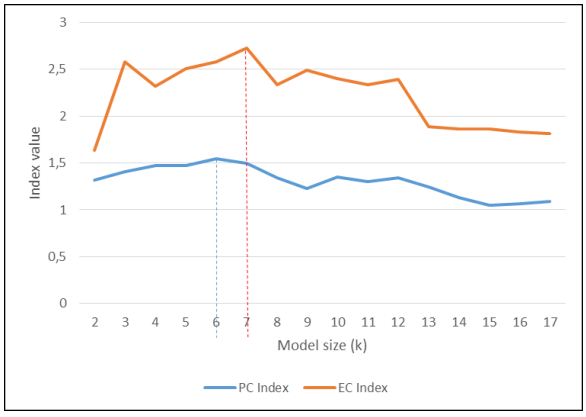

Fig. 4. Trends of PC and EC indexes on Reuters ZOO dataset (IGNGF clustering).

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. When there was dimension growth, these indexes were found to become 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]. Our PC index was found to perform slightly better than average but obviously 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 never found to fail in its estimation in our experimental context[8], whatever it faced with low or high dimensional estimation problem. Additionally, both 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 of the usal indexes namely the Xie-Beni (XI) index, and to a lesser extend, Calinski-Harabasz (CH) and

[8]Taking into account purity tolerance criteria due to imperfect behavior of clustering methods.

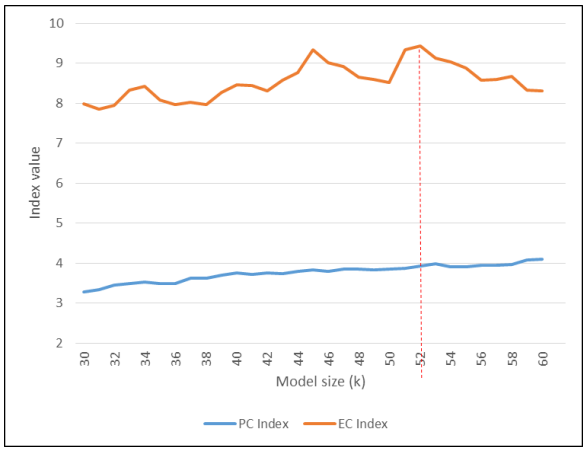

Fig. 5. Trends of PC and EC indexes on Reuters R52 dataset (IGNGF clustering).

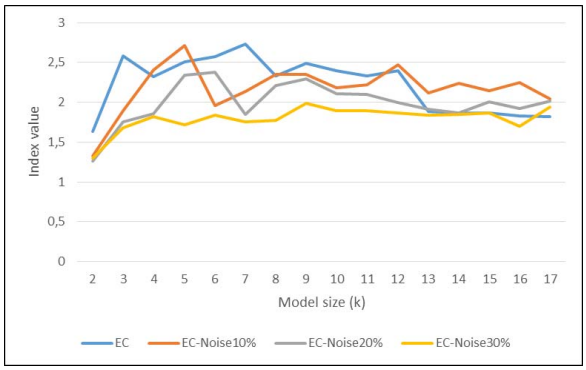

Fig. 6. Trends of EC indexes on UCI ZOO dataset with and without noise (IGNGF clustering).

Silhouette (SI) indexes.

The figure 4 and 5 show the evolution trends of the EC and PC indexes in the case of the ZOO and R52 datasets, respectively. On R52 related trends, it highlights the suitable index behavior (EC index) and in parallel the out-of-range index behavior mentioned earlier (PC index).

Interestingly, on the UCI ZOO dataset, the results of noise sensitivity analysis presented in table III underline the fact that noise has a relatively limited effect on the operation of PC and EC indexes. Figure 6 presents a parallel view of the different trends of EC value on non noisy and noisy clustering environments, respectively. It shows that noise tended to lower the index value in an overall way and soften the trends related to its behavior relative to changes in k value (see figure 3). However, the EC index was again found to have the most stable behavior in that context. As for the Silhouette index, this 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 indexes (DU) were found to shift from a correct to a wrong estimation as soon as noise began to appear.

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 2,2GHz quadricore processor and 8 GB of memory.

VI. Conclusion

We have proposed a new set of indexes for clustering quality evaluation relying on feature maximization measurement. This method exploits the information derived from features which could be associated to clusters by means of their associated data. Our experiments 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. 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.

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.

Acknowledgement

This work was done under the project ISTEX (http://www.istex.fr). ISTEX receives assistance from the French government and is managed by the National Research Agency under the program "Future Investments" bearing the reference ANR-10-IDEX-0004-12.

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., Guzmn-Arenas A., Godoy-Calderon S. and Barrn Fernndez 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, pp. 377-453 (1996)
[5] Calinsky, T. and Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics, 3(1):1-27 (1974)

[6] Davies, D. L. and Bouldin, D. W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2):224-227 (1979)
[7] Dimitriadou E., 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] Dunn, J.: Well separated clusters and optimal fuzzy partitions. Journal of Cybernetics, 4:95-104 (1974)
[9] Falk, I., Lamirel J.-C., Gardent C.: Classifying French Verbs Using French and English Lexical Resources. Proccedings of ACL, Jeju Island, Korea (2012)
[10] 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.), pp. 625-632 (1995)
[11] 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, pp. 703-715 (2012)
[12] 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)
[13] Guyon, I. and Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research, 3:1157-1182 (2003)
[14] Halkidi, M. and Batistakis, Y. and Vazirgiannis, M.: On clustering validation techniques, Journal of Intelligent Information Systems, 17(2):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), Innsbruck, Austria, pp. 97-103 (2008)
[17] Kolesnikov A., Trichina E. and Kauranne T.: Estimating the number of clusters in a numerical data set via quantization error modeling, Pattern Recognition, 48(3): 941952 (2015)
[18] Lago-Fernández 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 Mall, R. and Cuxac, P. and Safi, G.: Variations to incremental growing neural gas algorithm based on label maximization. Proceedings of IJCNN 2011, San Jose, CA, USA, pp. 956-965 (2011)
[20] Lamirel, J.C.: A new approach for automatizing the analysis of research topics dynamics: application to optoelectronics research. Scientometrics 93(1):151-166 (2012)
[21] 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, pp. 1-18 (2014)
[22] MacQueen, J. B.: Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, pp. 281-297 (1967)
[23] 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)
[24] 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)
[25] 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)
[26] Sun, L. and Korhonen, A. and Poibeau, T. and Messiant, C.: Investigating the cross-linguistic potential of VerbNet-style classification Proceedings of ACL, Beijing, China, pp. 1056-1064 (2010)
[27] 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, pp. 911-916 (2010)
[28] 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)