A Generic Methodology for Clustering to Maximises Inter-Cluster Inertia

A. Alaoui[1, 3]
[1]SIMO: Laboratory of Engineering Sciences and Modeling, Department of Physics
Faculty of Sciences, Ibn Tofail University
Kénitra, Morocco

B. Olengoba Ibara[2, 3]
[2]Laboratory of Ecology and Environment
Faculty of Sciences Ben M'sik, University Hassan II
Casablanca, Morocco

B. Ettaki[3, 1]
[3]Lyrica: Laboratory of Research in Computer Science, Data Sciences and Knowledge Engineering, Department of Data, Content and knowledge Engineering
School of Information Sciences
Rabat, Morocco

J. Zerouaoui[1]
[1]SIMO: Laboratory of Engineering Sciences and Modeling, Department of Physics
Faculty of Sciences, Ibn Tofail University
Kénitra, Morocco

***Abstract*—This paper proposes a novel clustering methodology which undeniably manages to offer results with a higher inter-cluster inertia for a better clustering. The advantage obtained with this methodology is due to an algorithm that showed beforehand its efficiency in clustering exercises, MC-DBSCAN, which is associated to an iterative process with a potential of auto-adjustment of the weights of the pertinent criteria that allows the reclassification of objects of the two closest clusters through each iteration, as well as the aptitude of the auto-evaluation of the precision of the clustering during the clustering process. This work conducts the experiments using the well-known benchmark, 'Seismic', 'Landform-Identification' and 'Image Segmentation', to compare the performance of the proposed methodology with other algorithms (K-means, EM, CURE and MC-DBSCAN). The experimental results demonstrate that the proposed solution has good quality of clustering results.**

Keywords*—*MC-DBSCAN; iterative process; inter-cluster inertia; unsupervised precision-recall metrics

I. Introduction

Nowadays, Data Mining [1] is imposed as one of the effective techniques for searching and retrieving information from very large databases. Like other search traditional operations, data mining is in the same vein. It aims to analyze a set of raw data in order to extract information that can be considered part of knowledge, and therefore, become exploitable. However, the data mining field specifically supplies solutions targeting the problematic of description, estimates, prediction, association, segmentation, classification and clustering [2]-[4]; To that end, the state of art shows that clustering and classification are both the most fundamental tasks in Data Mining.

The supervised classification always depends on a pre-constituted database reference. On the other hand, the exploitation of dataset without reference classification, unsupervised classification techniques called 'clustering' are unconditionally used [5]. For Clustering techniques [6], there is a choice among the methods based on the partition [7], hierarchical methods [8], [9], methods based on the grids [10], methods using models [11] and methods based on the density [12]. In that sense, Jain suggests, in his recent works, that 'There is no best clustering algorithm' [13], [14]. Furthermore, the practice shows that the performance of an algorithm depends on the tool choice and adaptation in accordance with the problem constraints.

The present paper proposes a generic methodology leading to an iterative process, that allows to improve in an optimal way the results of a clustering exercise. To that end, the density algorithm MC-DBSCAN [15] was used as the main clustering algorithm. This is justified by the fact that the MC DBSCAN showed its performance towards problems of multi-criteria in clustering. Specifically, in an earlier study, the MC-DBSCAN algorithm has given respectively the Accuracy values [17], [18] 93% and 34% with databases 'Vehicle-Silhouettes' and 'Iris' [15]. Although accuracy levels are high, some elements more or less misclassified are detected.

For this purpose, the performance of the solution proposed by this work will have as an assessment element for comparison, the results from the 'clustering' achieved with algorithms MC-DBSCAN [15], CURE [8], EM [19] and K-Means [20] each respectively representing a particular clustering category, clustering algorithms density, hierarchical clustering, clustering from clustering model and partitioning.

The outline of this article focuses successively on the presentation of MC-DBSCAN algorithm, the methodology governing the proposed solution; the treatment and comparing results obtained; and a conclusion.

The next part of this work, after the first section where the theme was introduced, is divided into five sections: Section 2, describing the original MC-DBSCAN algorithm; Section 3 presenting the proposed new generic methodology of clustering

in detail; Section 4, explaining the experimental results and discussions; Section 5, drawing conclusions.

II. MC-DBSCAN Algorithm

MC-DBSCAN is an improved version of DBSCAN [16] for the purpose of solving the problem of multi-criteria in clustering. The multi-criteria data is defined on different scale types with varied weights according to the importance of each criterion. This capacity has largely influenced the algorithm choice for the needs of this work, since MC-DBSCAN offers a possibility to adjust the weight of pertinent criteria in each iteration.

The MC-DBSCAN algorithm is composed of the following steps:

- Selection of an arbitrary object from a set of objects $alt_i \in D$;
- Calculation of similarity (Table 1/ function 1) and strong dissimilarity (Table 1/ function.3) of this object alt_i with each object from the set D;
- Calculation of weighted similarity (Table 1/ function 2) of this object alt_i;
- Calculation of overall similarity (Table 1/ function 4) of this object alt_i;
- The test of the value of overall similarity (Table 1/ function 4) and the presence of strong dissimilarity (Table 1/ function 3) allow the determination of the object which is considered to be a neighborhood of the object alt_i;
- The retrieval of each object density-connected to the object alt_i according to the parameters of overall similar (Table 1/ function 4) and the parameter $MinPts$:
- If alt_i is a core object, a cluster is formed;
- If alt_i is a point of border, therefore no point can be density-connected to alt_i and the algorithm visits the following object of the set D.

TABLE I. Functions of MC-DBSCAN Algorithm

Functions	**Meaning**
Similarity: $Similarite_i(alt_1, alt_2):$ $D \times D \to \{-1,0,1\}$	$Similarite_i(alt1, alt2) = \begin{cases} +1 \text{ si } \lvert alt1_i - alt2_i \rvert \le \sigma_i \\ -1 \text{ si } \lvert alt1_i - alt2_i \rvert \ge \sigma_i \end{cases}$
Weighted similarity: $SP(alt1, alt2):$ $D \times D \to [-1,1]$	$SP(alt1, alt2) = \sum_{i=1}^{n} p_i * Similarite_i(alt1, alt2)$
Strong dissimilarity: $DF(alt1, alt2):$ $D \times D \to \{0,1\}.$	$DF_i(alt1, alt2) = \begin{cases} 1 \, if \ \lvert alt1_i - alt2_i \rvert \ge \delta_i^+ \\ 0 \, elseif \end{cases}$
Similarities: $GS(alt1, alt2):$ $D \times D \to [-1,1]$	$GS(alt1, alt2) =$ $mm(SP(alt1, alt2), -DF_1(alt1, alt2), \ldots, -DF_m(alt1, alt2))$
Min and Max: $mm: [-1,1]^q \to [-1,1]$	$mm(p_1, \ldots, p_q) = \begin{cases} \max(p_1, \ldots, p_q) \, si \ p_i \ge 0 \\ \min(p_1, \ldots, p_q) \, si \ p_i \le 0 \\ 0 \sin on \end{cases}$

III. Proposed Methodology

The proposed solution is a generic methodology that can use other types of clustering algorithms. However, for the raised reasons in the previous parties, the MC-DBSCAN is proved to be the appropriate tool. In substance, the methodology is a model operating in an iterative manner to achieve the clustering. The iterative process of this model is tributary to the quality of the concluded clusters from the previous iteration. In other words, the process's continuity relies on the automatic comparison of the quality of the two consecutive iterations results. The solution consists of three principals steps.

The first phase leads to the MC-DBSCAN algorithm's intervention, which, first of all, uses the default values of inputs parameters for the preliminary classification. In this way, the obtained clusters serve as input data for the next stage, which is a procedure of the iterative classification.

The second phase represents the analysis and assessment stage of the obtained results in order to detect the similarity between the different achieved clusters. The analysis and assessment of the classification quality is done by calculation of the similarity between the clusters; in the sense that hence, the two clusters presenting a high similarity rate, show in contrast, an inter-class inertia [21] value less elevated (1). This situation would be a result of two possible scenarios, either the objects constituting two closest classes should belong to the same class, or an error is produced in the classification of certain objects that would belong normally to a class whereas they were found in the other class and vice versa.

The proposed model overcomes these classification anomalies by identifying the pertinent criteria (2), which would amplify the similarity between two classes, while taking back into consideration their weights in the following classification by using the AHP method [22], [23].

$$d(A,B) = \frac{p_A * p_B}{p_A + p_B} d_2(G_A, G_B)^2 \qquad (1)$$

p_A and p_B are respectively the weights of the two clusters A and B,

G_A and G_B are respectively the centers of the two clusters A and B.

$$R_{(ij,Criterion_k)} = \frac{I(C_i) + I(C_j)}{I(C_i, C_j)} \qquad (2)$$

$I(C_i)$ and $I(C_j)$ represents the respectively average distance between the elements and center of class *'i'* and class *'j'*.

$I(C_iC_i)$ represents the average distance between the two classes' centers *'i'* and *'j'*.

The third phase purpose is the evaluation of the two consecutive iterations. It concretely allows a comparison of the quality of the obtained results in the two last iterations in such a manner that the results' quality of the iteration *(i)* is better than the iteration *(i-1)*. In this case, the process of classification continues in order to improve the classes precision; if not, it restores and considers the issued results of the previous iteration *(i-1)* to complete the classification process.

For the purpose of assessing the overall quality of the results, the art of state offers several metrical approaches, which can be grouped into two categories. The first category are methods depending on the availability of a reference database. And the second category includes methods that do not use the reference database [24], this is namely inertial methods [21], Dunn [25], DB [26], Silhouette [27] and so on.

However, these preceding methods are limited in the evaluation of the results' quality in some clustering cases as mentioned in the work of Kassab [28].

To overcome this dilemma, Lamiel and other [29]-[32] have proposed improvements of the subsequent methods (Recall, Precision and F-Measures) based on reference classification, by making them adequate and relevant to unsupervised classification.

Nevertheless, the suggested method has been previously adapted for the clustering applied to text data.

However, the present paper proposes also the improvements of the following unsupervised methods: Recall, Precision and F-Measures, for being adaptable to all different types of data.

The principle of this work relies on the fact to be able to measure the classes' homogeneity by studying the distribution of intervals of each criterion within these classes. Consequently, each class is characterized by a set of intervals, in which the ratio of their weights inside the considered class and those in the partition should be maximal.

The global values of unsupervised Recall (4), Precision (5) and F-measure (6) are calculated as follows (Table 2):

$$Recall_{Unsupervised} = \frac{1}{|\overline{P}|}\sum_{C\in\overline{P}}\frac{1}{S_C}\sum_{Int_{i,j}\in S_C}\frac{|C_{Int_{i,j}}|}{|P_{Int_{i,j}}|} \quad (4)$$

$$Precision_{Unsupervised} = \frac{1}{|\overline{P}|}\sum_{C\in\overline{P}}\frac{1}{S_C}\sum_{Int_{i,j}\in S_C}\frac{|C_{Int_{i,j}}|}{|C|} \quad (5)$$

$$F1_{Unsupervised} = 2\cdot\left[\frac{1}{Recall_{Unsupervised}}+\frac{1}{Precision_{Unsupervised}}\right] \quad (6)$$

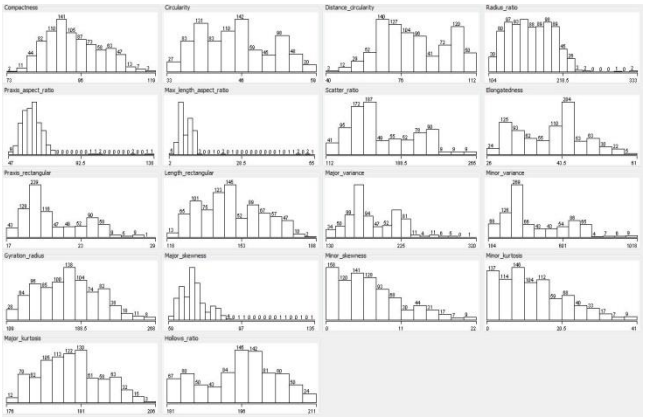

Fig. 1. Exemplary set of intervals of each criterion.

TABLE II. PROPORTIES OF UNSUPERVISED RECALL AND PRECISION

Properties	Meaning
Set of criteria that describe the data.	$E=\{Criterion_1, Criterion_2, \cdots, Criterion_m\}$ (7)
Values of each criterion are presented as a set of intervals (Fig. 1).	$Criterion_i=\{\text{Int}_{i,1}, Int_{i,2}, \cdots, Int_{i,k}\}$ (8)
Partition resulted from a clustering.	$P=\{C_1, C_2, \cdots, C_k\}$ (9)
S_c Represents all intervals of each criterion that have a maximal value of weight in class $C\in P$ in relation to other classes.	$S_C=\left\{\begin{matrix}\text{int}_{i,j}\in Criterion_i,\\ Criterion_i\in E\left\|\overline{W}_C^{\text{Int}_{i,j}}=\max(\overline{W}_{C'}^{\text{Int}_{i,j}})\right.\end{matrix}\right\}$ (10) $\overline{W}_C^{Int_{i,j}}=\frac{N_C}{N}$ (11)
N_c	Number of appearance of $Int_{i,j}$ within class C.
N	Number of appearance of $Int_{i,j}$ within the other classes.
$C_{Int_{i,j}}$	Set of objects of class C possessing the property $Int_{i,j}$.
$P_{Int_{i,j}}$	Set of objects of the partition P possessing the property $\text{Int}_{i,j}$.
$\overline{P}$: Set of proper classes.	$\overline{P}=\{C\in P\|S_c\neq\phi\}$ (12)

The following chart summarizes the process of the proposed methodology (Fig. 2).

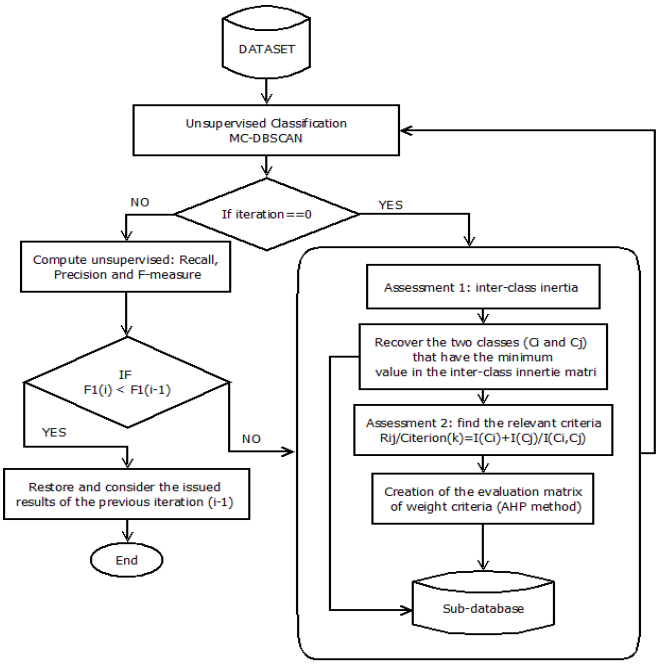

Fig. 2. Proposed methodology.

IV. RESULTS AND DISCUSSION

A. Databases Used

The performance of the proposed generic methodology and those of other algorithms namely EM, Cure, K-means and MC-DBSCAN are evaluated using the well-known reference databases, 'Seismic', 'LandformIdentification' and 'Image Segmentation' (Table 3). The three databases are from the great platform of data 'UCI Machine Learning Repository'.

TABLE III. DESCRIPTIONS OF DATASETS

Data Set	Instances	Criteria	Number of Classes
Seismic	2584	19	2
LandformIdentification	300	6	13
Image Segmentation	2310	19	6

B. Assessment Measures

To evaluate and compare the proposed methodology performance, we use the standard metrics: 'Precision: number of objects correctly assigned divided by total number of objects assigned', 'Recall: number of objects correctery assigned divided by the total number of objects that should be assigned' and 'F-measure: harmonic mean of precision and recall' which use the confusion matrix.

The precision scales the clusters in terms of the proportion of data that contain the specific properties of these first. Consequently, the more the data associated with a cluster have specific common properties, the more they are similar to each other, and therefore the criterion of homogeneity within the clusters is strengthened.

The Recall 'Recall' allows to measure the completeness of the clusters' contents, linked to the presence of specific properties that are specific to them. The more a cluster has a set of specific properties that are exclusive, the more it differs from other clusters, and therefore the criterion of heterogeneity between clusters is strengthened.

The F-measure which combines the precision and the recall is their harmonic average, named F-measure or F-score.

C. Results and Discussion

The Table 4 below includes and shows the results of different performed tests (Precision, Recall, F-measure) with the three test databases (a), (b) and (c). In these tests, the input parameters of the three first algorithms (EM, Cure and K-means) have default values except the parameter that represents the clusters' number which is fixed according to the issued information of reference databases.

TABLE IV. NUMERICAL RESULTS OF EM, CURE, K-MEANS, MC-DBSCAN AND THE PROPOSED METHODOLOGY

	Precision	Recall	F1
EM	75,52	76,19	75,85
CURE	53,36	57,5	55,35
K-MEANS	22,81	100	37,15
MC-DBSCAN	83,49	82,5	82,99
Proposed methodology	89,66	95,38	92,43

[a]With dataset: Seismic

	Precision	Recall	F1
EM	32,18	43,21	23,96
CURE	41,27	47,63	35,75
K-MEANS	30,83	45,22	21,02
MC-DBSCAN	63,54	38,90	48,26
Proposed methodology	84,73	83,68	84,20

[b]With dataset: Landform- Identification

	Precision	Recall	F1
EM	54,307	55,85	52,80
CURE	66,00	55,71	60,42
K-MEANS	58,62	54,27	56,36
MC-DBSCAN	73,35	62,58	67,54
Proposed methodology	83,53	81,98	82,75

[c]With dataset: Image Segmentation

The proposed 'Precision' factor for appreciating this work's results shows an important contrast between the achieved results of the propounded methodology and those of other existing algorithms. The suggested methodology presents respectively values 83,5%, 84% and 89% with databases 'Image Segmentation', 'Land form Identification' and 'Seismic'. In the other hands, the three other algorithms present fluctuating values between 23% and 83%, knowing that the number of clusters are pre- defined in these algorithms.

On one hand, these results lead to note that the precision's levels of achieved clusters are superior to 80% (required values for a sufficient homogeneity of clusters). This outcome illustrates or lets us foresee a high homogeneity within the given clusters from the proposed methodology. On the other hand, this methodology permitted an improvement of results of MC-DBSCAN algorithm with regard to its exclusive use. It allowed an improvement of the of clusters homogeneity varying between 6% and 21% in accordance to the used 'test databases'.

Regarding the 'Recall' factor, the suggested methodology gives an average of 87% for the three test databases. However, it points out respectively the average values of 61%, 66%, 53% and 57% for MC-DBSCAN, K-means, Cure and EM algorithms. Exceptionally, in the third database 'Seismic', the value of the 'Recall' factor, issued from K-means algorithm, has shown the existence of clusters that present a set of specific properties that are exclusive for them. This means that The 'Recall' shows a value of 100% (against 95% for the proposed methodology).

Moreover, the improvement provided by the proposed methodology is important and considerable. It is 26% in comparison to the result given by the MC-DBSCAN algorithm. This improvement emanates from the inclusion of iterative corrections, which allow a re-classification of misclassified items in previous iterations.

Overall, the harmonic average of the two factors 'Precision' and 'Recall' on the three test databases has demonstrated an improvement respectively of 20%, 48%, 35% and 36% compared to the MC-DBSCAN, K- means, Cure and EM algorithms, which highlights the relevance and pertinence of the proposed methodology.

V. CONCLUSION

Due to the recurring difficulty that rises in the evaluation of the quality of a clustering, many approaches are used in the performance estimation in a clustering exercise results. The state of art puts forward approaches of appreciation based on

the judgment of an expert, the use of the labeled data when available, the comparison with the references classification or the computation of various indices generally relying on the relations of intra\extra distances clusters. Even though those approaches offered results that are relatively satisfying in some projects, it still reveals its limits in certain clustering exercises. However, the proposed methodology seems to be an alternative solution to overcome the limitations faced with the approaches mentioned above. The methodology leading to an iterative process, that allows to improve in an optimal way the results of a clustering exercise with a higher inter-cluster inertia. To that end, MC-DBSCAN algorithm was used as the main clustering algorithm.

As a minimum, it would be important to mention that this methodology highlighted the improvement of the inter-class inertia; nevertheless, in order to achieve a better precision of clusters, it is better and significant to include a parallel evaluation, which would allow an optimized intra-cluster quality and a better homogeneity.

In addition, the proposed methodology could also contribute, beyond the MC-DBSCAN algorithm, to the improvement of the performance and to the precision of other multi-criterion assistance with the decision algorithms, as long as it offers the possibility to adjust the weights of the criteria's from iteration to the other.

REFERENCES

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases". AI magazine, 17(3) , 37, 1996.

[2] J. Han and M. Kamber, "Data Minin Concepts and Techniques", Morgan Kaufman, San Francisco,CA, 2001.

[3] D. Hand, H. Mannila, and P. Smith, "Principles of Data Mining", MIT Press, Combridge, MA, 2001.

[4] Marcoulides, and A. George, "Discovering Knowledge in Data: an Introduction to Data Mining", Journal of the American Statistical Association, doi: 10.1198/jasa.2005.

[5] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data clustering: A review, ACM Computing Surveys", 31, 264-323, 1999.

[6] K. Koutroumbas and S. Theodoridis, "Pattern Recognition, Academic Press, 2009.

[7] L. Kaufman, and P.J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis", Wiley-Interscience, New York, second ed, 2005.

[8] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient Clustering algorithm for large databases", Proc. Of ACG SIGMOD Intl. Conf. on Management of Data, pp. 73-82, 1998.

[9] M. Chis, D. Dumitrescu, "Evolutionary Hierarchical Clustering for Data Mining, Proceedings of the Symposium Zilele Academice Clujene", Computer Science Section, Seminar on Computer Science, pp.12-18, 14-22 June 2002.

[10] R. Agrawal et al. "Automatic subspace clustering of high dimensional data for data mining applications", Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, ACM Press (1908), pp.94–105, 1998.

[11] C. Fraley, and A.E. Raftery, "Model-based clustering, discriminant analysis, and density estimation", Journal of the American Statistical Association, pp.611–631, 2002.

[12] A. Hinneburg, and D.A. Keim, "An efficient approach to clustering in large multimedia databases with noise", Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, pp. 58–65, 1998.

[13] J. Kleinberg, "An impossibility theorem for clustering", In: NIPS 15. pp. 463– 470, 2002.

[14] K.J. Anil, "Data Clustering: 50 Years Beyond K-Means", the 19th International Conference on Pattern Recognition (ICPR), Tempa, FL, December 8, 2008.

[15] A. Idrissi, and A. Alaoui, "A Multi-Criteria Decision Method in the DBSCAN Algorithm for Better Clustering" International Journal of Advanced Computer Science and Applications(IJACSA), 7(2), 2016. http://dx.doi.org/10.14569/IJACSA, 2016.

[16] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD96) 226-231, 1996.

[17] Powers and M.W. David, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation", Journal of Machine Learning Technologies, Vol.2, No.1, pp.37–63, 2011.

[18] Q. Xiao and E.G. McPherson, "Tree health mapping with multispectral remote sensing data at UC Davis", California. Urban Ecosystems 8: 349-361, 2005.

[19] Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm". Journal of the Royal Statistical Society, 39(1) :1–38, 1977.

[20] F. Husson, J. Josse, and J. Pagès,"Principal component methods - hierarchical clustering partitional clustering: why would we need to choose for visualizing data? ", Technical report, 2010.

[21] L. Lebart, A. Maurineau, M. Piron, "Traitement des données statistiques", Dunod, Paris, 1982.

[22] T.L. Saaty, "The Analytic Hierarchy Process", McGrawHill, 1980.

[23] T.L. Saaty, "Decision-making with the AHP: Why is the principal Eigenvector necessary", European Journal of Operational Research 145 -85–91, 2003.

[24] M. Ghribi, P. Cuxac, J.C. Lamirel, and A. Lelu, "Mesures de qualité de clustering de documents : Prise en compte de la distribution des mots-clés", Atelier EvalECD, Hamamet, Tunisie, 2010.

[25] J. Dunn, "Well Separated clusters and optimal fuzzy partitions", Journal of Cybernetics, 4, 95-104, 1974.

[26] D.L. Davies and D.W. Bouldin, "A cluster separation measure", IEEE Trans. Pattern Anal. Machine Intell, 1(4), 224-227, 2000.

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

[28] R. Kassab, J.-C. Lamirel, "A Multi-level Abstraction Model for Competitive Learning", Artificial Intelligence and Applications - AIA 2008, 97-103, 2008.

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

[30] J.C. Lamirel, Z. Boulila, M. Ghribi, P. Cuxac, "A new incremental growing neural gas algorithm based on clusters labeling maximization: application to clustering of heterogeneous textual data", Twenty Third International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE 2010), 1-4 june, Cordoba, Spain, 2010.

[31] D.D. Lewis, Y. Yang, T. Rose, and F. Li, "RCV1: A New Benchmark Collection for Text Categorization Research", Journal of Machine Learning Research, 5:361-397, 2004.

[32] P. Cuxac, A. Lelu, M. Cadot, "Incremental follow-up of evolutions in an indexed information base: an evaluation / correction loop for the choice of algorithms and parameters", ["Suivi incrémental des évolutions dans une base d'information indexée: une boucle évaluation/correction pour le choix des algorithmes et des paramètres", 2ème conférence Internationale sur les systèmes d'informations et Intelligence Economique SIIE 2009, Hammamet Tunisie, 2009].