Exploitation d'une mesure de micro-précision cumulée non supervisée pour l'évaluation fiable de la qualité des résultats de clustering

Jean-Charles Lamirel, Maha Ghribi, Pascal Cuxac

EXPLOITATION D'UNE MESURE DE MICRO-PRECISION CUMULEE NON SUPERVISEE POUR L'EVALUATION FIABLE DE LA QUALITE DES RESULTATS DE CLUSTERING

Jean-Charles Lamirel*, Maha Ghribi**, Pascal Cuxac**
*LORIA - Campus Scientifique BP 239 - 54506 Vandœuvre-lès-Nancy, France
**INIST-CNRS, 2 allée du Parc de Brabois, 54500-Vandœuvre-lès-Nancy, France

Résumé
Dans le cadre de la classification non supervisée, ou clustering, le fait de ne pas disposer d'une classification de référence représente un lourd handicap pour évaluer la performance des algorithmes. De leur coté, les critères traditionnels de qualité (inertie, DB…) ne permettent pas d'estimer correctement la qualité du clustering dans de nombreux cas, comme dans celui des données textuelles. Nous présentons ainsi une approche alternative pour l'évaluation de la qualité du clustering basée sur des critères de rappel, de précision et de F-mesure non supervisés, exploitant les descripteurs des données associées aux clusters. Le rappel permet de mesurer l'exhaustivité du contenu des clusters en termes de descripteurs propres, spécifiques à chaque cluster. La précision mesure l'homogénéité des clusters en termes de proportion des données contenant leurs descripteurs propres associés. Ce papier se concentre particulièrement sur la construction d'un nouvel index de micro-précision cumulée qui permet d'évaluer la qualité globale d'un résultat de clustering tout en donnant la possibilité complémentaire de distinguer clairement entre des résultats homogènes et des résultats hétérogènes. La comparaison expérimentale du comportement des critères classiques avec notre nouvelle approche est effectuée sur un ensemble de références bibliographiques issues de la base de données de PASCAL.

1 Introduction

En classification non supervisée (clustering), le fait de ne pas avoir de classification de référence sur laquelle s'appuyer représente un lourd handicap pour évaluer la performance des algorithmes. Il existe certes des indices de qualité basés sur des calculs de distance dont les plus connus sont l'inertie inter-classes et intra-classes (Lebart et al., 1982) :

- L'inertie Intra - classes permet de mesurer le degré d'homogénéité entre les données associées à une classe. Elle calcule leurs distances par rapport au point représentant le profil de la classe.

$$Intra = \frac{1}{n}\sum_{C\in P}\frac{1}{2n_c}\sum_{i\in C}\sum_{j\in C} d(i,j)^2 \quad (1)$$

- L'inertie Inter - classes mesure le degré d'hétérogénéité entre les classes. Elle calcule les distances entre les points représentant les profils des différentes classes de la partition.

$$Inter = \frac{1}{n}\sum_{C\in P} n_C d^2(c, c_G) \quad (2)$$

où c est le centre de la classe C et c_G est le centre du nuage de points matérialisant les données.

Parmi les autres Indices de qualité utilisant la distance entre individus qui ont été développés, on citera l'indice de Dunn (Dunn, 1974), l'indice de validation de Davies-Bouldin (Davies et Bouldin, 2000) et la Silhouette (Rousseeuw, 1987). Pour plus d'informations sur ces indices on se reportera à (Ghribi et al., 2010). Tous ces indices, ne permettent cependant pas d'estimer la qualité du

clustering dans bon nombre de cas, comme dans celui des données textuelles (Kassab et al., 2008). Nous avons donc développé une approche alternative basée sur des mesures de Rappel, Précision et F-mesure non supervisée exploitant les descripteurs des données associées aux classes (Lamirel et al., 2004). Les premiers tests sur ces mesures nous ont conduits à en proposer de nouvelles adaptations que nous présentons ci-après.

2 Rappel – Précision non supervisés

Dans le domaine de la recherche d'information, le Rappel *R* représente le rapport entre le nombre de documents pertinents restitués pour une requête donnée et le nombre total de documents pertinents qui auraient été trouvés dans la base de données documentaire. La Précision *P* représente le rapport entre le nombre de documents pertinents qui ont été restitués pour une requête donnée et le nombre total de documents retournés pour ladite requête. Ces indices ont été adaptés par (Lamirel et al., 2004) au cas de la classification non supervisée, les mesures ne se faisant plus sur des documents mais sur les descripteurs des données associées aux classes, ou clusters, à l'issue du processus de clustering :

Soit une partition $P = (C_1;\ldots;C_k)$ issue d'une classification non supervisée d'un ensemble de documents. On définit alors les Macro Rappel- Précision comme les valeurs moyennes de Rappel et de Précision pour chaque cluster. Ils prennent les formes suivantes :

$$R_M = \frac{1}{\left|\overline{P}\right|}\sum_{C\in\overline{P}}\frac{1}{S_C}\sum_{p\in S_C}\frac{\left|c_p\right|}{\left|P_p\right|}\ ;\quad P_M = \frac{1}{\left|\overline{P}\right|}\sum_{C\in\overline{P}}\frac{1}{S_C}\sum_{p\in S_C}\frac{\left|c_p\right|}{\left|c\right|} \tag{3}$$

où c_p présente l'ensemble des données du cluster *C* possédant la propriété *p* et P_p représente l'ensemble des données de la partition *P* possédant la propriété *p.*

S_c représente l'ensemble des propriétés propres au cluster *C* :

$$S_C = \left\{ p\in d, d\in C_i\, C \;\middle|\; \overline{W}_C^p = \max_{C'\in P}(\overline{W}_{c'}^p)\right\},\ \text{avec}\ \overline{W}_C^p = \frac{\sum_{d\in C} W_d^p}{\sum_{C'\in P}\sum_{d\in C'} W_d^p} \tag{4}$$

où W_p^d représente le poids de la propriété *p* pour un document *d*,

et $\overline{W}_C^p$ représente le rapport du poids cumulé de la propriété *p* dans le cluster *C* par rapport à son poids total dans la partition. On définit l'ensemble des clusters propres de la partition *P* comme suit :

$$\overline{P} = \left\{ C\in P \middle| S_C \neq \varnothing \right\} \tag{5}$$

Le Rappel permet de mesurer l'exhaustivité du contenu des clusters, lié à la présence de propriétés propres qui leur sont spécifiques. Plus un cluster présente un ensemble de propriétés propres qui lui sont exclusives, plus il se distingue des autres clusters, et donc plus le critère d'hétérogénéité entre clusters est renforcé.

La Précision mesure l'homogénéité des clusters en termes de proportion de données contenant les propriétés propres de ces premiers. Plus les données associées à un cluster présentent des propriétés propres communes, plus elles sont similaires entre elles, et donc plus le critère d'homogénéité à l'intérieur des clusters est renforcé.

Le Macro-Rappel et la Macro-Précision ont des comportements inverses en fonction du nombre de classes. Ainsi, ces indices permettent d'estimer de manière globale un nombre optimal de clusters pour une méthode donnée et pour un ensemble de données fixé. La meilleure partition est dans ce cas celle qui minimise l'écart entre leur valeur.

Il a été montré par (Lamirel et al., 2004) qu'un des avantages déterminants de cette approche, qui s'inspire de l'analyse du comportement des classificateurs symboliques, est d'être indépendante de la méthode de clustering utilisée, contrairement aux approches basées sur la distance. Elle permet donc de comparer les méthodes entre elles. Cependant, son défaut principal est que la Macro-Précision, en particulier, est peu sensible à la présence de clusters hétérogènes de fort effectif, surtout dans le cas de l'existence conjointe d'un nombre important de clusters de faible taille (Ghribi et al., 2010).

Pour corriger cela, nous définissons ci-après de nouveaux indices de Micro – Rappel/Précision, calculés en moyennant directement les valeurs de Rappel/Précision sur l'ensemble des propriétés propres, et non plus sur les clusters :

$$R_m = \frac{1}{|L|} \sum_{c \in \overline{C}, p \in S_c} \frac{|c_p|}{|P_p|}; \; P_m = \frac{1}{|L|} \sum_{c \in \overline{C}, p \in S_c} \frac{|c_p|}{|c|} \tag{6}$$

où L représente la dimension de l'espace de description.

Les Micro- Rappel/Précision possèdent des caractéristiques générales analogues aux Macro-Rappel/Précision. Cependant, en les mixant avec ces derniers indices, ils permettent d'identifier des résultats de clustering hétérogènes. En effet, dans ce dernier cas, les Précisions des clusters de petite taille ne compenseront plus celles des clusters de grande taille et les propriétés imprécises présentes dans ces derniers, s'ils s'avèrent hétérogènes, auront un effet considérable sur la Micro-Précision. Par conséquent, même si la Macro- et la Micro-Précision mesurent toutes deux le degré d'homogénéité des clusters, l'écart entre ces deux mesures permet de confirmer la présence de clusters hétérogènes de taille importante.

Il est cependant possible de se reporter uniquement aux indications fournies par les indices de Micro-Précision et de Micro-Rappel, si le calcul de la Micro-Précision est opéré de manière cumulée. Dans ce dernier cas, l'idée directrice est celle de donner une influence prépondérante à la Micro-Précision relative aux clusters de taille les plus importantes, étant donné que ceux-ci sont plus spécialement susceptibles de rapatrier de l'information hétérogène, et donc, à eux seuls, de faire baisser la qualité des résultats de clustering. Ce calcul peut être opéré comme suit :

$$PC_m = \sum_{i=|c_{\inf}|,|c_{\sup}|} \frac{1}{\left|\overline{C_{i+}}\right|^2} \sum_{c \in \overline{C_{i+}}, p \in S_c} \frac{|c_p|}{|c|} \Big/ \sum_{i=|c_{\inf}|,|c_{\sup}|} \frac{1}{\left|\overline{C_{i+}}\right|} \tag{7}$$

où C_{i+} représente le sous-ensemble des clusters de C pour lesquels le nombre de données associées est supérieur à i, et

$$\inf = \arg\min_{c_i \in C} |c_i|, \sup = \arg\max_{c_i \in C} |c_i| \tag{8}$$

Une F-mesure cumulée qui combine la Micro-Précision cumulée avec le Micro-Rappel peut être construite de la manière suivante :

$$FC_m = \frac{2(R_m \times PC_m)}{R_m + PC_m} \tag{9}$$

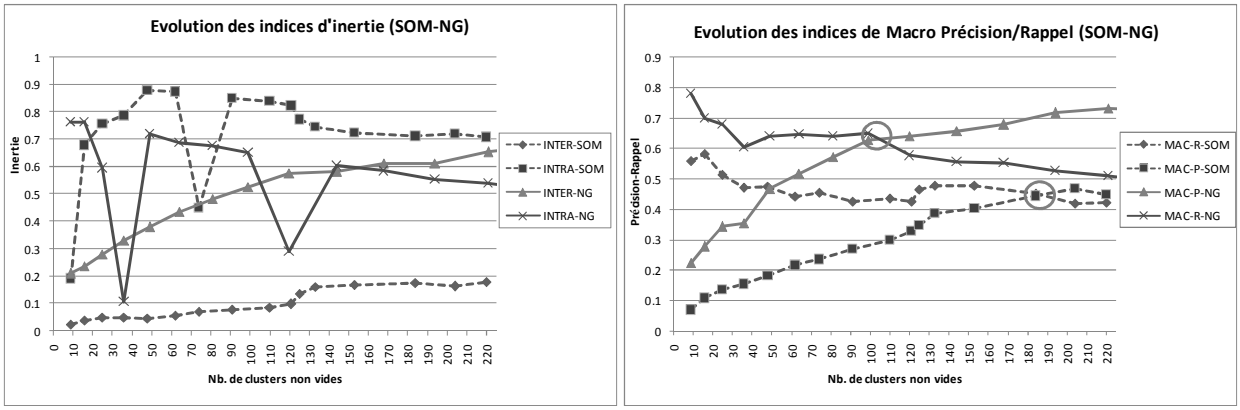
Un facteur de pénalisation peut également être appliqué à la méthode de clustering si celle-ci produit des classes vides. Il vient alors une nouvelle F-mesure cumulée corrigée qui peut s'exprimer comme :

$$FCC_m = \frac{\overline{|C|}}{|C|} \times FC_m \tag{10}$$

Comme nous le montrons dans l'expérience décrite ci-après, cette dernière technique permet également d'exploiter l'indice de F-mesure cumulée corrigée pour la détection d'un nombre optimal de clusters.

3 Expérimentation

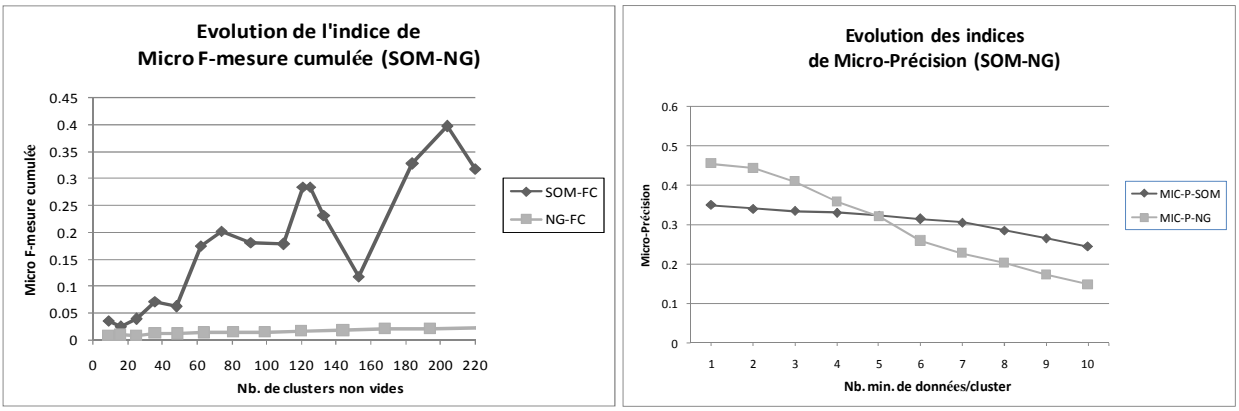
Pour illustrer le comportement de nos nouveaux indices, nous avons utilisé un corpus test de 1341 documents sur le thème de la recherche en Lorraine, issus de la base de données PASCAL du CNRS. Ces documents sont indexés par 889 descripteurs de fréquence supérieure ou égale à 3. Pour effectuer le clustering nous avons exploité parallèlement les méthodes SOM (Kohonen, 1982) et NG (Martinetz et al. 1994). L'analyse des résultats effectuée par un expert a montré que seule la méthode SOM fournissait des clusters homogènes sur ce corpus. Les résultats présentés à la figure 1A illustrent le fait que les indices classiques d'inertie ont un comportement instable qui ne leur permet pas d'identifier clairement un nombre optimal de clusters, que ce soit dans le contexte de SOM ou dans celui de NG. D'un autre coté, il apparait également, à la figure 1B, que le comportement des indices de Macro- Rappel/Précision est stable et qu'il permet d'identifier un nombre optimal de clusters dans tous les cas. En effet, en première approximation, ce nombre optimal peut être obtenu au point de croisement des courbes de Macro-Rappel et de Macro-Précision (c.à.d. 100 clusters pour NG et 256 clusters pour SOM sur la figure 1B). Toutefois, aucun des groupes d'indices précédemment cités ne permet d'estimer correctement la qualité des résultats de clustering. En effet, aucun de ceux-ci ne permet de discriminer entre des résultats de clustering homogènes (SOM) et des résultats hétérogènes (NG). Dans les deux cas, ils présentent même le défaut important de privilégier ce dernier type de résultats.



FIGURES 1A ET 1B – *Evolution des valeurs des indices d'inertie (1A) et de Macro-Précision et de Macro-Rappel (1B) en fonction du nombre de clusters non vides pour les méthodes SOM et NG).*

Une estimation adéquate des résultats de clustering peut par contre être obtenue par l'exploitation des résultats fournis par l'indice de Micro-F-mesure cumulée (Eq. 9-10), lui-même directement dérivé de l'indice de Micro-Précision cumulée (Eq. 7). Dans le cas de NG, la valeur de Micro-F-mesure cumulée, qu'elle soit corrigée (Eq. 10), ou non (Eq. 9), reste très faible, ceci quelque soit le nombre de clusters considéré (Figure 2A). Ce phénomène est principalement dû à l'influence de la Micro-Précision des clusters de taille importante. Cela traduit le fait que les propriétés propres des clusters dans les partitions générées par NG sont largement moins précises que celles des clusters

produits par SOM. L'évolution des courbes de Micro-précision standard (Eq. 6) en fonction de la taille des clusters permet également de vérifier que ce phénomène touche plus particulièrement les clusters volumineux dans le cas de NG (Figure 2A). De plus, quelque soit la méthode étudiée, l'indice de Micro-F-mesure cumulée permet d'assurer un suivi précis de la qualité en fonction des configurations choisies. Dans le cas de SOM, la perte de qualité apparaissant pour certaines tailles de grille (par ex. au point 160 sur la figure 2A) qui induisent la formation de gros clusters hétérogènes est en effet directement caractérisée par la chute des valeurs de cet indice. La figure 3 illustre finalement l'intérêt de l'indice de Micro-F-mesure cumulée corrigée pour la détection du nombre optimal de clusters. La courbe associée à cet indice présente un plateau qui permet de situer le maximum d'efficacité du clustering. Il apparait que celui-ci se situe aux environs du point de croisement des courbes de Macro-Rappel et de Macro-Précision (Figure 1A). Cela semble donc corrélativement entériner le choix de ce point de croisement comme point caractéristique des résultats de clustering.



FIGURES 2A ET 2B – *Evolution des valeurs des indices de Micro-F-mesure cumulée (FC) relativement au nombre de clusters non vides (2A) de Micro-Précision standard (MIC-P) relativement à leur taille (2B) pour les méthodes SOM et NG.*

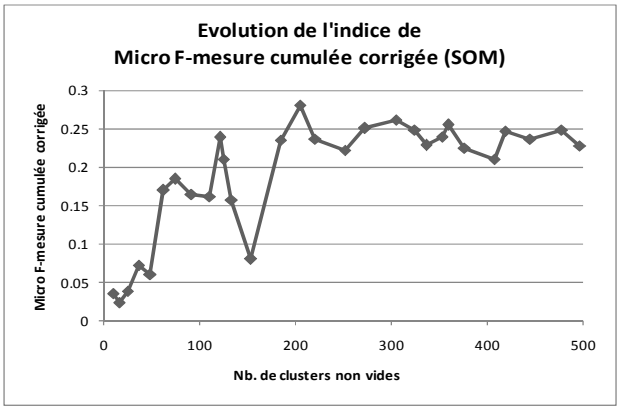


FIGURE 3 – *Evolution des valeurs des indices de Micro-F-mesure cumulée corrigée (FCC) relativement au nombre de clusters non vides pour la méthode SOM.*

4 Conclusion

Nous avons proposé une nouvelle approche pour l'évaluation de la qualité du clustering basée sur l'exploitation des propriétés associées aux classes par l'intermédiaire de l'indice Micro-Précision cumulée non supervisée et de ses extensions. Nous avons montré, que contrairement aux indices classiques, nos nouveaux indices permettaient d'évaluer précisément la qualité globale d'un résultat de clustering tout en donnant la possibilité complémentaire de distinguer clairement entre des résultats homogènes et des résultats hétérogènes. Nous avons également montré que cette approche s'appliquait aussi bien à la comparaison des résultats fournis par des méthodes différentes, qu'à

l'analyse précise des résultats fournis par une méthode donnée, en fonction de paramètres d'entrée tels que le nombre de clusters. Même si des tests complémentaires restent naturellement à réaliser, il apparait déjà clairement que cette approche représente une avancée significative dans le sens d'une meilleure analyse, et donc d'une meilleure exploitation, des résultats de clustering. De plus, sa généricité permet d'envisager de l'appliquer directement à des résultats d'analyse impliquant tout type de corpus de données, quelque que soit sa nature, textuelle ou autre.

Bibliographie

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

[2] Dunn J. (1974): Well Separated clusters and optimal fuzzy partitions. Journal of Cybernetics,4, 95-104.

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

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

[5] Kohonen T. (2001): Self-Organising Maps. 3rd ed., Springer-Verlag, Berlin.

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

[7] Lebart L.,Maurineau A., Piron M. (1982) : Traitement des données statistiques. Dunod, Paris.

[8] Martinetz T., Schulten K.(1994): Topology representing networks. Neural Network., 7(3), 507-522.

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

Abstract

In the context of unsupervised classification, or clustering, the fact of not having a reference classification represents a heavy handicap to evaluate the performance of the algorithms. On their own side, traditional quality indexes (Inertia, DB…) do not allow to properly estimate the quality of the clustering in several cases, as in that one of the textual data. We thus present an alternative approach for clustering quality evaluation based on unsupervised measures of Recall, Precision and F-measure exploiting the descriptors of the data associated with the obtained clusters. The Recall makes it possible to measure the exhaustiveness of the contents of the clusters in terms of peculiar descriptors specific to each cluster. The Precision measures the homogeneity of the clusters in terms of proportion of data containing the associated peculiar descriptors. This paper especially focuses on the construction of a new cumulative Micro precision index that makes it possible to evaluate the overall quality of a clustering result while clearly distinguishing between homogeneous and heterogeneous results. The experimental comparison of the behavior of the classical indexes with our new approach is performed on a dataset of bibliographical references issued from the PASCAL database.