Les méthodes de classification non supervisées appliquées aux textes : mesure de la performance des résultats de clustering de documents

Pascal Cuxac, Jean-Charles Lamirel, Maha Ghribi

Pascal Cuxac
INIST / CNRS, Vandoeuvre-lès-Nancy, France
Jean-Charles Lamirel
LORIA, Vandoeuvre-lès-Nancy, France
Maha Ghribi
INIST / CNRS, Vandoeuvre-lès-Nancy, France

Les méthodes de classification non supervisées appliquées aux textes : mesure de la performance des résultats de clustering de documents

Résumé : Nous présentons une approche alternative pour l'évaluation de la qualité de classifications non supervisées de textes basée sur des critères de rappel, précision et F-mesure non supervisés, exploitant les descripteurs associées aux classes. La comparaison expérimentale du comportement des critères classiques avec notre approche est effectuée sur des données bibliographiques.

Abstract: This paper presents an alternative approach to measuring the quality of non-supervised text classification based on the recall, precision and non-supervised F-measure criteria, using class descriptors. The experimental comparison of classical criteria behaviour to our approach is based on bibliographic data.

1. Introduction

L'utilisation des méthodes de classification de l'information est devenue courante pour analyser de gros corpus de données que ce soit par exemple dans le cadre de besoins de veille scientifique ou d'analyses stratégiques de la recherche. Elles s'appuient sur la combinaison de techniques de classification automatique (supervisée ou non) et de cartographie pour visualiser le résultat.

En procédant à une classification, on cherche à construire des ensembles homogènes d'individus, c'est-à-dire partageant un certain nombre de caractéristiques identiques. En outre, le clustering (ou classification non supervisée) permet de mettre en évidence ces regroupements sans connaissance a priori sur les données traitées. Si celles-ci sont des publications scientifiques (données textuelles), comme dans le cas que nous traitons, et que l'on considère le corpus de départ comme représentatif d'un domaine, les classes peuvent être assimilées à des thèmes de recherches de ce domaine. La classification obtenue, représentée sous forme de carte, permet alors d'avoir une vue d'ensemble du domaine scientifique traité.

Il existe de nombreuses méthodes de clustering que Turenne (Turenne 2001) regroupe en 8 familles. Dans nos expérimentations nous exploitons des méthodes issues de plusieurs de celles-ci, comme les méthodes neuronales (SOM, NG…), les méthodes de partitionnement (K-means…) et les méthodes basées sur des graphes (Walktrap, Germen). Certaines de ces méthodes sont bien connues (SOM, K-means…), d'autres ont été développées par nous-mêmes comme Germen (Lelu et al. 2006) et I²GNG (Lamirel et al. 2010).

Un problème central que l'on doit alors se poser est de qualifier les résultats obtenus en termes de qualité : un indice de qualité est un critère qui permet en effet de décider quelle méthode utiliser, de fixer un nombre de classes optimal, ou encore, d'évaluer et/ou mettre au point une nouvelle méthode.

2. Mesure de la performance des résultats de clustering de documents

En classification non supervisée et avec nos types de données il est très difficile d'avoir une classification de référence pour évaluer la performance des algorithmes. On pourrait envisager d'utiliser un corpus de référence étiqueté tel que celui des dépêches de l'agence 'Reuters' qui est régulièrement exploité dans les campagnes d'évaluation, comme dans celles des systèmes de recommandation. La difficulté est alors de faire l'indexation de ces textes et d'évaluer ensuite la part d'erreur due à la méthode d'indexation automatique de celle due à l'algorithme de clustering utilisé et à son paramétrage. Dans le cas du corpus 'Reuters', notre expérience montre que l'indexation proposée par (Lewis et al. 2004) est finalement peu adaptée à l'évaluation de méthodes de classifications non supervisées de données textuelles (Cuxac et al. 2009).

Procéder sans classification de référence consiste alors à se tourner vers l'exploitation d'indices de qualité utilisables en mode non supervisé. Un état de l'art du domaine permet de recenser des indices basés sur des calculs de distance : on citera entre autres l'inertie inter-classes et intra-classes (Lebart et al. 1982), l'indice de Dunn (Dunn 1974), l'indice de Davies–Bouldin (Davies et Bouldin 2000) et la Silhouette (Rousseeuw 1987). Du fait de l'exploitation possible de liens entre les documents ou leurs mots-clés, l'utilisation d'indices opérant sur des graphes, comme par exemple la modularité (Newman et Girman 2004) ou la performance (Van Dongen 2000), est également envisageable.

Nos expérimentations ont cependant montré qu'aucun de ces indices ne permettait d'estimer correctement la qualité d'un résultat de clustering sur des données textuelles. Ceux-ci ne permettent notamment pas de discriminer entre des résultats de classification homogènes et des résultats hétérogènes. Ils peuvent même présenter le défaut important de privilégier cette dernière famille de résultats (Ghribi et al. 2010).

Partant de ces constatations, et en nous inspirant du comportement des classificateurs symboliques, nous avons développé de nouveaux indices de qualité uniquement basés sur l'exploitation des propriétés associées aux classes, à savoir les indices de Rappel, Précision et de F-mesure non supervisés (Lamirel et al. 2004).

Le Rappel permet de mesurer l'exhaustivité du contenu des classes, 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 classes en termes de proportion de données contenant les propriétés propres de ces premières. 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. On définit les Macro Rappel-Précision comme les valeurs moyennes de Rappel et de Précision pour chaque classe. Ces indices permettent d'estimer de manière globale un nombre optimal de classes 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 démontré (Lamirel et al. 2004) que

notre approche présentait l'avantage déterminant, relativement à celles basées sur la distance, d'être totalement indépendante de la méthode de clustering utilisée, et donc qu'elle permettait de comparer objectivement les méthodes de classification entre elles. Cependant, nos travaux récents ont également montré, qu'en présence de grosses classes hétérogènes, la Macro-Précision n'était pas assez discriminante (Ghribi et al. 2010). Nous proposons donc ici une extension de notre approche basée sur la définition 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 classes.

Les expérimentations que nous présentons illustrent l'application de nos indices de qualités étendus aux résultats de clustering obtenus avec plusieurs méthodes de différentes familles sur un corpus issu d'une base de données bibliographique de référence (Base PASCAL, www.inist.fr), sur le thème de la recherche en Lorraine (1341 documents indexés par 889 descripteurs de fréquence supérieure ou égale à 3). Pour plus de clarté nous ne présenterons ici que les résultats obtenus avec les méthodes SOM (Kohonen, 1982) et NG (Martinetz et al. 1994). La figure 1a montre que, avec nos données et les méthodes testées, les indices classiques d'inertie ont un comportement instable qui ne leur permet pas d'identifier clairement un nombre optimal de clusters. Le comportement de nos indices de Macro-Rappel/Précision est stable et permet d'identifier un nombre optimal de clusters dans tous les cas (Fig. 1b)

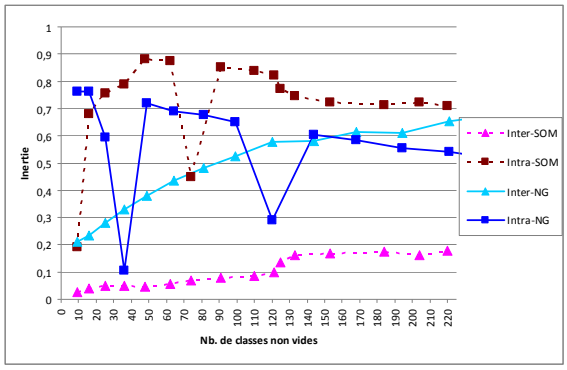

Fig. 1a : Evolution des indices d'inertie (SOM-NG)

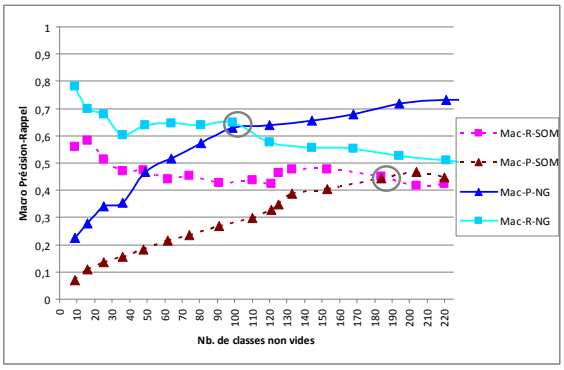

Fig. 1b : Evolution des indices de Macro Précision-Rappel (SOM-NG)

Dans le cas de NG, les différences entre les valeurs de Micro et de Macro-Précision sont toujours plus importantes que dans le cas de SOM, quelque soit le nombre de clusters considéré (Fig. 2a). 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 en fonction de la taille des clusters permet de vérifier que ce phénomène touche plus particulièrement les clusters volumineux (Fig. 2b).

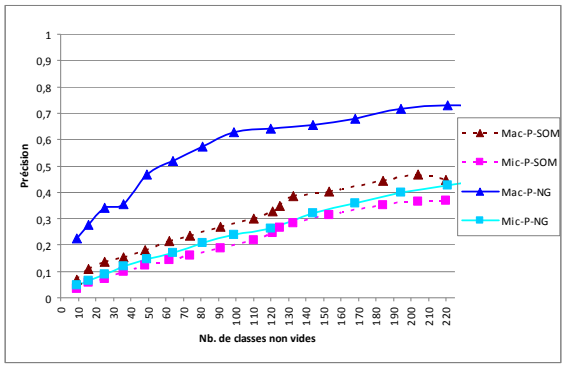

Fig. 2a : Evolution des indices de Macro/Micro Précision (SOM-NG)

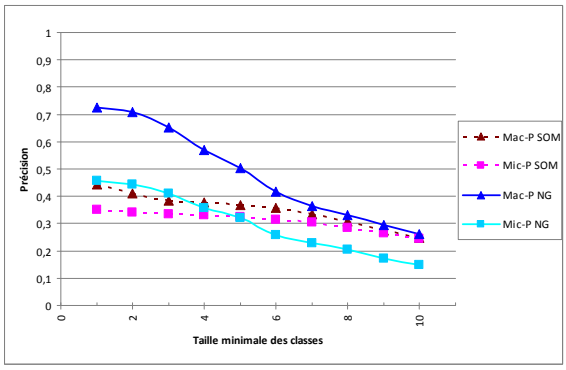

Fig. 2b : Evolution des indices de Macro/Micro Précision en fonction de la taille des classes (SOM-NG)

3. Conclusion

Nos expérimentations démontrent, que contrairement aux indices de qualité de la littérature, nos indices étendus, indépendant de la méthode de clustering utilisée, permettent de distinguer clairement les méthodes fournissant des résultats de classification homogènes de celles fournissant des résultats hétérogènes sur un même corpus.

Ces indices permettant d'évaluer la qualité globale d'une classification non supervisée demandent cependant d'analyser en détail leurs courbes d'évolution, rendant leur utilisation complexe pour un utilisateur non averti. Nous devons maintenant nous intéresser à la visualisation de ces résultats afin d'estimer rapidement et très facilement la qualité d'un résultat de clustering de textes.

4. Bibliographie

Cuxac P., Lelu A., Cadot M. 2009. 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.

Davies D.L., Bouldin D.W. 2000. A cluster separation measure. IEEE Trans. Pattern Anal. Machine Intell, 1(4), 224-22.

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

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.

Kohonen T. 1982. Self-organized formation of topologically correct feature maps. Biological Cybernetics, vol. 43, pp 56-59.

Lamirel J.C., Boulila Z., Ghribi M., Cuxac P. 2010. 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.

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.

Lebart L., Maurineau A., Piron M. 1982. Traitement des données statistiques. Dunod, Paris.

Lelu A., Cuxac P., Johansson J. (2006). Classification dynamique d'un flux documentaire: une évaluation statique préalable de l'algorithme GERMEN. JADT 2006, Besançon, 19-21 Avril 2006, p. 617-630.

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

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

Newman M.E.J., Girman M. 2004. Finding an evaluating community structure in networks. Physical Review E, 69(6).

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.

Turenne N. 2001. Etat de l'art de la classification automatique pour l'acquisition de connaissances à partir de textes, 30p. Technical Report Inra 2001.

Van Dongen S.M. 2000. Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht