Évaluation en cascade d'algorithmes de clustering

Laurent Candillier[*,**], Isabelle Tellier[*], Fabien Torre[*], Olivier Bousquet[**]

[*]GRAppA - Université Charles de Gaulle - Lille 3
[**]Pertinence - 32 rue des Jeûneurs - 75002 Paris

L'évaluation des résultats d'algorithmes de clustering, ainsi que la comparaison de tels algorithmes, reste encore aujourd'hui une problématique ouverte importante. La difficulté vient principalement du fait que de telles évaluations sont subjectives par nature car il existe souvent différentes manières pertinentes de regrouper un même ensemble de données.

Les techniques existantes dans ce cadre souffrent de quelques limitations. Utiliser des données artificielles ou un expert d'un domaine particulier ne permet pas de généraliser les résultats à différents types de données réelles. Comparer les clusters obtenus aux classes prédéfinies d'un jeu de données étiqueté est rarement approprié car d'autres regroupements peuvent être plus pertinents. Utiliser des critères numériques tels l'inertie intra-cluster et/ou la séparation inter-clusters est également subjectif par nature car basé sur des notions prédéfinies de la pertinence d'un clustering (or parfois, des clusters qui se chevauchent sont plus pertinents que des clusters séparés).

En fait, ce que l'on souhaite évaluer, c'est la capacité d'une méthode de clustering à fournir de la connaissance nouvelle et utile dans un certain cadre. L'idée principale de notre approche consiste à considérer le clustering comme un prétraitement à une tâche que l'on sait évaluer : l'apprentissage supervisé par exemple. Ainsi, si les résultats d'un algorithme supervisé sont améliorés lorsqu'il est aidé par de l'information provenant d'un algorithme de clustering, alors nous postulons que cela signifie que le clustering a fourni une information nouvelle et utile.

Adaptant la technique de *cascade generalization* [Gama et Brazdil (2000)] au cas où un apprenant est non supervisé, la méthode d'*évaluation en cascade* d'algorithmes de clustering que nous proposons consiste donc à :

1. effectuer un apprentissage supervisé sur un jeu de données étiqueté,
2. effectuer un apprentissage supervisé sur le même jeu de données enrichi par les résultats de l'algorithme de clustering évalué,
3. et comparer les erreurs des deux classifieurs appris.

Cette méthode nous permet alors d'évaluer objectivement l'intérêt de l'information capturée par le clustering. De plus, la diminution du taux d'erreur de l'algorithme supervisé lorsqu'il est aidé par l'information issue du clustering nous permet de quantifier cet intérêt.

Nous avons mené plusieurs expérimentations avec cette méthode, en considérant :

1. plusieurs jeux de données issus de l'UCI Machine Learning Repository [Blake et Merz (1998)],
2. différentes façons d'enrichir les jeux de données à partir des résultats d'algorithmes de clustering, parmi lesquelles celle proposée dans [Apte et al. (2002)],

3. et différents algorithmes supervisés : C4.5 [Quinlan (1993)] basé sur la construction d'arbres de décision, C5 boosté [Quinlan (2004)] basé sur la combinaison de plusieurs arbres de décision, et DLG [Webb et Agar (1992)] basé sur l'utilisation de *moindres généralisés*.

Nous avons ainsi évalué plusieurs algorithmes de clustering utilisant des modèles de complexités différentes : depuis K-means jusqu'à certains modèles probabilistes plus complexes.

Sur chaque jeu de données, pour chaque méthode d'enrichissement, chaque algorithme supervisé et chaque algorithme de clustering utilisés, nous faisons cinq validations croisées avec découpage du jeu de données en deux, comme proposé dans [Dietterich (1998)]. Pour chaque validation croisée, nous calculons les taux d'erreur pondérés de l'algorithme supervisé avec ou sans information ajoutée par l'algorithme de clustering évalué. Puis nous utilisons quatre mesures pour comparer les résultats :

1. le nombre de victoires de chaque méthode sur l'ensemble des jeux de données considérés,
2. le nombre de victoires significatives, en utilisant le *5×2cv F-test* [Alpaydin (1999)] pour vérifier si les résultats sont significativement différents,
3. le *wilcoxon signed rank test*, qui indique si une méthode est significativement meilleure qu'une autre sur un ensemble de problèmes indépendants,
4. et l'erreur pondérée moyenne.

Nous avons ainsi mis en avant que quels que soient la méthode d'enrichissement et l'algorithme supervisé utilisés, l'ordre dans lequel les méthodes de clustering sont rangées par notre méthode d'évaluation reste toujours le même. Plus particulièrement, celle-ci met en avant le fait que plus le modèle utilisé par l'algorithme de clustering est complexe, plus l'information qu'il capture est pertinente, en cela qu'il aide l'algorithme supervisé à améliorer ses performances. Ce résultat, loin de surprendre, montre le comportement cohérent de notre méthode.

Références

Alpaydin, E. (1999). Combined 5x2cv F-test for comparing supervised classification learning algorithms. *Neural Computation 11*(8), 1885–1892.

Apte, C. V., R. Natarajan, E. P. D. Pednault, et F. A. Tipu (2002). A probabilistic estimaton framework for predictive model analytics. *IBM Systems Journal 41*(3).

Blake, C. et C. Merz (1998). UCI repository of machine learning databases [http ://www.ics.uci.edu/∼mlearn/MLRepository.html].

Dietterich, T. G. (1998). Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation 10*(7), 1895–1923.

Gama, J. et P. Brazdil (2000). Cascade generalization. *Machine Learning 41*(3), 315–343.

Quinlan, J. R. (1993). *C4.5 : Programs for Machine Learning*. KAUFM.

Quinlan, R. (2004). Data mining tools see5 and c5.0.

Webb, G. I. et J. W. M. Agar (1992). Inducing diagnostic rules for glomerular disease with the DLG machine learning algorithm. *Artificial Intelligence in Medicine 4*, 419–430.