Comparaison de critères de pureté pour l'intégration de connaissances en clustering semi-supervisé

Germain Forestier, Cédric Wemmert, Pierre Gançarski

Université de Strasbourg - LSIIT - CNRS - UMR 7005
Pôle API, Bd Sébastien Brant - 67412 Illkirch, France
{forestier,wemmert,gancarski}@unistra.fr

Résumé. L'utilisation de connaissances pour améliorer les processus de fouille de données a mobilisé un important effort de recherche ces dernières années. Il est cependant souvent difficile de formaliser ce type de connaissances, comme celles-ci sont souvent dépendantes du domaine. Dans cet article, nous nous intéressons à l'intégration de connaissances sous la forme d'objets étiquetés dans les algorithmes de clustering. Plusieurs critères permettant d'évaluer la pureté des clusters sont présentés et leur comportement est comparé sur des jeux de données artificiels. Les avantages et les inconvénients de chaque critère sont analysés pour aider l'utilisateur à faire un choix.

1 Introduction

L'intégration de connaissances dans les algorithmes de *clustering* a connu un fort intérêt ces dernières années, des connaissances dites du domaine (*background knowledge*) étant souvent disponibles. Celles-ci peuvent se présenter sous des formes très différentes et sont difficiles à formaliser de manière générique car elles dépendent souvent du domaine d'application. Plusieurs travaux (Wagstaff et al., 2001; Bilenko et al., 2004) se sont intéressés à l'utilisation de connaissances sous la forme de contraintes entre paires d'objets. À l'instar d'un algorithme supervisé qui va apprendre une fonction de classification, cette information peut être utilisée pendant le processus de clustering pour guider l'algorithme vers une solution en accord avec ces connaissances.

Un concept récurrent dans les méthodes utilisant ce type de connaissances est la pureté des clusters qui consiste à évaluer la qualité des clusters par rapport à ces objets étiquetés. Un cluster pur sera un cluster dans lequel tous les objets, dont la classe est connue, appartiennent à une et une seule classe. Un cluster impur présentera des objets de classes différentes.

L'objectif de cet article est de présenter et de comparer différentes méthodes d'évaluation de la pureté des clusters. Dans la section 2, nous présentons un état de l'art de l'utilisation de connaissances en clustering. Dans la section 3, différentes mesures de pureté sont présentées et comparées. Enfin, la section 4 présente les conclusions et les futures pistes de recherche de ces travaux.

2 Évaluation d'un clustering

Nous considérons ici que la connaissance est un ensemble d'objets étiquetés. Soit N le nombre d'objets étiquetés, $\mathbb{C} = \{c_1, c_2, \ldots, c_K\}$ les clusters trouvés par l'algorithme de clustering, $\mathbb{W} = \{w_1, w_2, \ldots, w_C\}$ les classes des objets étiquetés, c_k les objets appartenant au cluster k et w_k l'ensemble des objets de la classe k, $|c_k|$ le nombre d'objets du cluster k et $n_j^i = |w_i \cap c_j|$ les objets à la fois dans le cluster i et de la classe j.

2.1 Calcul de pureté

La façon la plus simple de calculer la pureté est de chercher la classe majoritaire dans chacun des clusters et de sommer le nombre d'objets de cette classe pour chacun des clusters (Manning et al., 2008). La pureté d'un clustering se définit comme :

$$\mathbf{\Pi}_{\text{simple}}(\mathbb{C}, \mathbb{W}) = \frac{1}{N} \sum_{i}^{K} \arg\max_{j}(n_j^i) \tag{1}$$

Une autre façon de calculer la pureté des clusters est proposée par Solomonoff et al. (1998) qui le formulent comme la probabilité, étant donné un cluster, que deux objets tirés au hasard sans remise soient de la même classe :

$$\mathbf{\Pi}_{\text{prob}}(\mathbb{C}, \mathbb{W}) = \frac{1}{N} \sum_{i}^{K} |c_k| \pi_{\text{prob}}(c_i) \text{ avec } \pi_{\text{prob}}(c_i) = \sum_{j}^{C} \left(\frac{n_j^i}{|c_i|} \right)^2 \tag{2}$$

Cette mesure a l'avantage, par rapport à la pureté simple (Eq. 1), de prendre en compte la distribution des classes minoritaires d'un cluster (i.e. les classes autres que la classe majoritaire), et favorise donc les clusters présentant un nombre limité de classes différentes.

Ces deux mesures de pureté ont cependant un inconvénient majeur, qui est de surévaluer la qualité d'un clustering avec un nombre important de clusters. Différentes propositions ont été faites pour résoudre ce problème. Par exemple, Ajmera et al. (2002) ont proposé de calculer la pureté des clusters en terme de classes ainsi que la pureté des classes en terme de clusters, c'est à dire pour chaque classe sa répartition dans les différents clusters. Ces deux valeurs sont ensuite combinées pour donner une évaluation globale du clustering. Considérer également la pureté des classes permet de pénaliser un nombre trop important de clusters. La pureté des classes se calcule de manière similaire à la pureté des clusters mais en observant la distribution des clusters des objets dans chaque classe :

$$\mathbf{\Pi}_{\widetilde{\text{prob}}}(\mathbb{C}, \mathbb{W}) = \frac{1}{N} \sum_{i}^{K} |c_k| \pi_{\widetilde{\text{prob}}}(w_i) \text{ avec } \pi_{\widetilde{\text{prob}}}(w_i) = \sum_{j}^{C} \left(\frac{n_j^i}{|w_i|} \right)^2 \tag{3}$$

La pureté des clusters et la pureté des classes sont ensuite combinées :

$$\mathbf{\Pi}_{\text{overall}}(\mathbb{C}, \mathbb{W}) = \sqrt{\mathbf{\Pi}_{\text{prob}}(\mathbb{C}, \mathbb{W}) \times \mathbf{\Pi}_{\widetilde{\text{prob}}}(\mathbb{C}, \mathbb{W})} \tag{4}$$

Une autre approche consiste à considérer également une mesure de la qualité du clustering à partir des données. Demiriz et al. (1999) utilisent un algorithme pour optimiser la pureté des

clusters appelé Gini et similaire à Eq. 2. Pour éviter que l'algorithme ne génère une solution avec un nombre trop important de clusters, la fonction objective est une moyenne arithmétique de la pureté des clusters et de la qualité du clustering. La combinaison de ces deux critères permet d'éviter des solutions trop extrêmes.

Enfin, Eick et al. (2004) ont également proposé d'introduire une notion de pénalité par rapport au nombre de clusters de la solution proposée afin de résoudre ce problème. Cette méthode permet de pénaliser une solution ayant un nombre de clusters trop important par rapport au nombre de classes.

$$\text{penalty}(K) = \begin{cases} \sqrt{\frac{K-C}{N}} & \text{si } K \geq C \\ 0 & \text{sinon} \end{cases} \tag{5}$$

avec K le nombre de clusters, C le nombre de classes et N le nombre d'objets. Cette pénalité est retranchée de l'indice de pureté choisi, pondérée par un paramètre β, comme suit :

$$\mathbf{\Pi}_{\text{penalty}}(\mathbb{C}, \mathbb{W}) = \mathbf{\Pi}_{\text{simple}}(\mathbb{C}, \mathbb{W}) - \beta\text{penalty}(K) \tag{6}$$

Une autre solution est d'utiliser le cadre de la théorie de l'information et d'évaluer l'information mutuelle normalisée entre les connaissances et le clustering :

$$\mathbf{\Pi}_{\text{nmi}}(\mathbb{C}, \mathbb{W}) = \frac{I(\mathbb{C}, \mathbb{W})}{[H(\mathbb{C}) + H(\mathbb{W})]/2} \tag{7}$$

I est l'information mutuelle :

$$I(\mathbb{C}, \mathbb{W}) = \sum_i \sum_j \frac{n_j^i}{N} \log \frac{n_j^i / N}{|c_i|/N \times |w_j|/N} \tag{8}$$

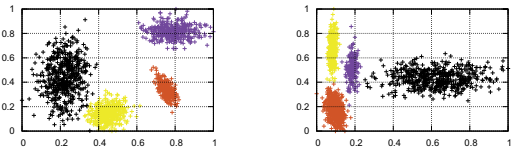
$$= \sum_i \sum_j \frac{n_j^i}{N} \log \frac{n_j^i \times N}{|c_i| \times |w_j|} \tag{9}$$

H est l'entropie :

$$H(\mathbb{W}) = -\sum_k \frac{|w_k|}{N} \log \frac{|w_k|}{N} \tag{10}$$

2.2 Comparaison de partitions

Un autre indice couramment utilisé est l'indice de *Rand* (Rand, 1971) qui permet de comparer des partitions. Dans notre cas, il consiste à vérifier si les couples d'objets de la même classe d'après les connaissances disponibles, ont été placés dans un même cluster. On dit qu'un couple d'objets est un vrai positif (VP) si les deux objets sont de la même classe et sont placés dans le même cluster, et un vrai négatif (VN) quand les deux objets sont de classes différentes et sont placés dans deux clusters différents. Un faux positif (FP) correspond à deux objets de



(a) (b)

FIG. 1 – *Deux jeux de données utilisés.*

classes différentes placés dans le même cluster. Un faux négatif (FN) correspond à deux objets de la même classe dans deux clusters différents. Il peut être défini de la manière suivante :

$$\mathbf{\Pi}_{\text{rand}}(\mathbb{C}, \mathbb{W}) = \frac{VP + VN}{VP + FP + FN + VN} \tag{11}$$

$(VP + FP + FN + VN)$ représentant tous les couples possibles d'objets et $(VP + VN)$ les couples d'objets correctement classés. L'indice de $Rand$ donne cependant un poids égal aux faux positifs et aux faux négatifs.

La *F-Mesure* (van Rijsbergen, 1979) quant à elle, permet de pondérer ces deux valeurs en tenant compte de la précision (P) et du rappel (R) :

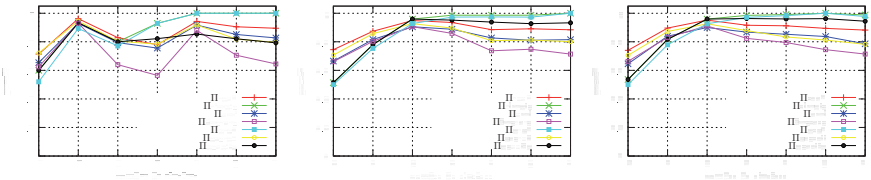
$$P = \frac{VP}{VP + FP} \quad R = \frac{VP}{VP + FN}$$

$$\mathbf{\Pi}_{\text{fmesure}}(\mathbb{C}, \mathbb{W}) = \frac{(\beta^2 + 1)P \times R}{\beta^2 P + R} \tag{12}$$

Si $\beta = 1$, valeur retenue dans cet article, la précision et le rappel ont alors la même importance. L'avantage de ces deux indices ($\mathbf{\Pi}_{\text{rand}}$ et $\mathbf{\Pi}_{\text{fmesure}}$) est qu'ils intègrent implicitement le nombre de clusters, en défavorisant naturellement les solutions avec un nombre de clusters trop important.

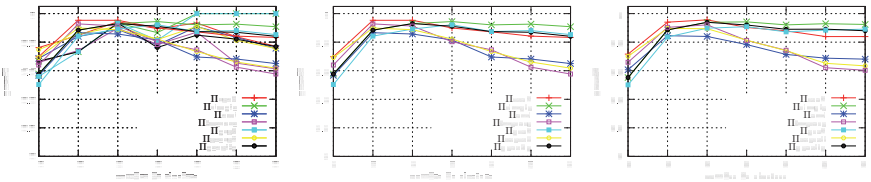
2.3 Évaluation des différents critères de qualité

Dans cette section nous allons évaluer les critères présentés dans la section précédente sur différents jeux de données de test. La figure 1 présente trois jeux de données artificiels représentant chacun quatre clusters dans un espace à deux dimensions. L'algorithme KMEANS a été utilisé sur ces jeux de données avec des nombres de clusters variant de 2 à 8 (8 étant deux fois le nombre de clusters attendu). Pour chaque clustering, les mesures présentées dans les sections précédentes ont été calculées. Trois configurations différentes ont été évaluées, la première avec 1% des données étiquetées, la seconde avec 10% des données étiquetées et enfin la dernière avec 25% des données étiquetées. Chaque expérience a été effectuée 100 fois avec des initialisations aléatoires de l'algorithme, puis les résultats ont été moyennés. Les figures 2 et 3 représentent les résultats respectivement pour les jeux de données des figures 1(a) et 1(b).



(a) 1% (b) 10% (c) 25%

FIG. 2 – *Evolution des critères en fonction du nombre de clusters pour le jeu donné Fig. 2 (a).*



(a) 1% (b) 10% (c) 25%

FIG. 3 – *Evolution des critères en fonction du nombre de clusters pour le jeu donné Fig. 2 (b).*

Quand le nombre d'objets étiquetés disponibles est faible (1%), la majorité des critères ont un comportement très aléatoire. En effet, il n'est pas du tout garanti d'avoir des objets pour chacune des classes du jeu de données. C'est pourquoi ces critères sont difficilement utilisables quand vraiment très peu de connaissances sont disponibles. Quand le nombre d'objets étiquetés augmente (10%), il est plus probable d'avoir des objets étiquetés pour chaque classe. Par conséquent, les courbes deviennent plus caractéristiques. Le problème présenté précédemment sur le fait que les mesures de pureté simples surévaluent la qualité du clustering quand le nombre d'objets augmente peut être observé. En effet, la pureté simple ($\mathbf{\Pi}_{\text{simple}}$) ainsi que la pureté par cluster ($\mathbf{\Pi}_{\text{prob}}$) ne font qu'augmenter avec le nombre de clusters. Les autres indices ($\mathbf{\Pi}_{\text{rand}}, \mathbf{\Pi}_{\text{nmi}}, \mathbf{\Pi}_{\text{fmesure}}, \mathbf{\Pi}_{\text{overall}}, \mathbf{\Pi}_{\text{penalty}}$) ont tendance à diminuer avec l'augmentation du nombre de clusters. Les plus caractérisques étant $\mathbf{\Pi}_{\text{fmesure}}$, $\mathbf{\Pi}_{\text{overall}}$ et le $\mathbf{\Pi}_{\text{nmi}}$. Les critères $\mathbf{\Pi}_{\text{rand}}$ et $\mathbf{\Pi}_{\text{penalty}}$ diminuent de façon moins caractéristique. Il est intéressant de noter qu'il n'y a pas de différence importante entre les résultats obtenus avec 10% d'objets étiquetés et 25% d'objets étiquetés.

3 Conclusion

L'intégration de connaissances dans les algorithmes de classification non supervisée représente un enjeu important. De plus en plus de connaissances implicites ou explicites sont disponibles et il convient de développer des approches permettant d'en tirer parti. Dans cet article, nous avons abordé l'utilisation d'objets étiquetés pour évaluer la pureté d'un résultat de clustering. Plusieurs critères ont été présentés, formalisés et comparés. Il en ressort que les critères éval-

uant la pureté sans prendre en compte le nombre de clusters peuvent rapidement surévaluer la qualité des résultats. Pour résoudre ce problème, il est possible de prendre en compte une mesure qui va pénaliser les résultats avec un nombre de clusters trop important. D'autres types de critères qui comparent le regroupement de couples d'objets prennent en compte implicitement le nombre de clusters. C'est notamment le cas de la *F-Mesure* qui a donné des résultats particulièrement bons lors de nos expériences.

Références

Ajmera, J., H. Bourlard, I. Lapidot, et I. McCowan (2002). Unknown-multiple speaker clustering using hmm. In *International Conference on Spoken Language Processing*, pp. 573–576.

Bilenko, M., S. Basu, et R. J. Mooney (2004). Integrating constraints and metric learning in semi-supervised clustering. In *International Conference on Machine Learning*, pp. 81–88.

Demiriz, A., K. Bennett, et M. Embrechts (1999). Semi-supervised clustering using genetic algorithms. In *Intelligent Engineering Systems Through Artificial Neural Networks*, pp. 809–814.

Eick, C. F., N. Zeidat, , et Z. Zhao (2004). Supervised clustering - algorithms and benefits. In *International Conference on Tools with Artificial Intelligence*, pp. 774–776.

Manning, C. D., P. Raghavan, et H. Schütze (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*, 622–626.

Solomonoff, A., A. Mielke, M. Schmidt, et H. Gish (1998). Clustering speakers by their voices. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, Volume 2, pp. 757–760.

van Rijsbergen, C. J. (1979). *Information Retrieval*. London : Butterworths.

Wagstaff, K., C. Cardie, S. Rogers, et S. Schroedl (2001). Constrained k-means clustering with background knowledge. In *International Conference on Machine Learning*, pp. 557–584.

Summary

In recent years, the use of background knowledge to improve the data mining process has been intensively studied. Indeed, background knowledge along with knowledge directly or indirectly provided by the user are often available. However, it is often difficult to formalize this kind of knowledge, as it is often dependent of the domain. In this article, we studied the integration of knowledge as labeled objects in clustering algorithm. Several criteria allowing the evaluation of the purity of a clustering are presented and their behavior is compared using artificial datasets. Advantages and drawbacks of each criteria are analyzed in order to help the user to make a choice among them.