

Analyse des variations entre partitions générées par différentes techniques de classification automatique de textes

Jean-François Chartier, Jean-Guy Meunier, Choukri Djellali

LANCI – UQAM - C.P. 8888, Succ. Centre-Ville, Montréal, Québec, Canada, H3C 2P8

Résumé

La classification automatique est une technique d'exploration et d'analyse de texte largement utilisée par la communauté des sciences humaines et sociales. Plusieurs dizaines d'algorithmes ont été conçus, chacun s'appuyant sur un principe d'induction différent. La comparaison du comportement de ces différents algorithmes est devenue un enjeu méthodologique fondamental. Trois stratégies de comparaison ont été suggérées, basées sur des critères externes, internes ou relatifs. Nous présentons les premiers résultats d'une étude basée sur une stratégie de comparaison relative. Cette étude a pour objectif de mesurer la variation dans les résultats de classification automatique de texte en fonction de l'algorithme utilisé. Quatre algorithmes sont comparés : les réseaux de neurones adaptatifs (ART1), le K-Means (KM), Expectation Maximisation (EM) et les cartes topologiques auto-organisatrices (SOM).

Abstract

Clustering algorithm is a technique of exploration and analysis of textual data widely used in social sciences and humanities. Several algorithms were conceived, each leaning on a different inductive principle. The comparison of the behaviour of these algorithms became an important methodological object of interest. Three strategies of comparison were suggested, based on external, internal or relative criteria. We present the preliminary results of a study based on a strategy of relative comparison criteria. The objective of this study is to measure the variation in the text clustering results, according to different algorithms. Four algorithms are compared: the adaptive resonance theory (ART1), the k-means (KM), expectation maximisation (EM) and self-organizing maps (SOM).

Keywords: text clustering, consensus analysis between partitions, methodology, ART1, K-means, SOM, EM

1. Contexte : les techniques de classification automatique de textes en sciences humaines et sociales

Les chercheurs des sciences humaines et sociales intègrent de plus en plus dans leurs pratiques de recherche, les techniques d'analyse issues de la fouille de texte (*Text Mining*). La classification automatique de textes, dite avec apprentissage « non-supervisé » aussi appelée « clustering », s'est révélée particulièrement intéressante dans nombre de programmes de recherche. Ces techniques ont été utilisées en philosophie (Meunier et al., 2005), en sociologie (Demazière et al., 2006), en littérature (Reinert, 1993), en sémiotique (Rastier, 2001) et les « Humanités digitales » (Hockey, 2001). Elles ont été utilisées notamment pour l'analyse thématique (Forest and Meunier, 2004), l'analyse des mondes lexicaux (Reinert, 1993), l'analyse conceptuelle (Meunier and Forest, 2009), la visualisation des données (Lebart, 2004), l'analyse des attracteurs

culturels (Gagnon, 2004), l'analyse des représentations sociales (Lalhou, 2003; Kalampalikis and Moscovici, 2005), l'analyse des réseaux sociaux (Diesner and Carley, 2005) et l'analyse documentaire en général (Salton, 1989). Ces différents chercheurs appliquent des techniques de classification automatique sur des corpus de textes très variés, composés d'articles de presse, de pages web, de verbatims, d'œuvres littéraires ou philosophiques, etc.

1.1. Problématique et questions de recherche

La classification automatique peut être définie en termes logiques par un triplet (O, X, G) où est défini un ensemble d'objets $O = \{o_1, \dots, o_n\}$; un ensemble de variables $X = \{x_1, \dots, x_n\}$ décrivant chaque objet de O ; et G un principe d'induction quelconque. À partir de ces informations, une opération de classification est définie ainsi: pour tous les objets de l'ensemble O , $(G(x_1, \dots, x_n))$.

Appliquée à la classification textuelle, cette définition cependant cache une méthodologie complexe qui implique plusieurs choix d'opérationnalisation : les objets soumis à la classification sont identifiés par la segmentation d'un corpus soit par phrases, par paragraphes ou un autre type de segment de texte; les variables en fonction desquelles sont classés les objets sont sélectionnées soit par l'indexation des lexèmes, des lemmes, des n-grammes ou autres types d'unité textuelle; et le principe d'induction en fonction duquel sont générées les classes implique la sélection d'un algorithme de classification. Ce dernier, en particulier, doit se faire parmi une variété toujours grandissante d'algorithmes, qui peuvent être de nature statistique, par partition itérative ou hiérarchique, neuronale statique ou dynamique, probabiliste, génétique, rigide ou floue, ou même mixte. On en dénombre plus d'une centaine de différents dans la littérature (Jain et al., 1999; Berkhin, 2006) et les principes d'induction formalisés par ces algorithmes sont très différents les uns des autres (Estivill-Castro, 2002).

Dans la communauté des sciences humaines et sociales, cette méthodologie et les choix d'opérationnalisation qui lui sont liés sont parfois occultés par l'utilisation d'un logiciel commercial « clé en main », comme ALCESTE, WORDSTAT, SPSS, POLYANALYST, Le SPHINX et bien d'autres. Ainsi, le choix de l'un ou l'autre des algorithmes de classification n'est pratiquement jamais discuté, les conséquences sur la partition (les classes) qui sera découverte, sont sous-estimées. Les différences formelles entre algorithmes soulèvent pourtant des questions importantes : est-ce que celles-ci entraînent également des différences empiriques significatives ? Autrement dit, sur un même ensemble d'objets (de segments de texte), est-ce que deux chercheurs qui appliquent une classification automatique chacune à l'aide d'un algorithme différent vont se retrouver avec des partitions également différentes ? Si c'est le cas, quelle est l'ampleur de la variation ? Si celle-ci est grande, en quoi cela questionne la confiance que ces chercheurs peuvent avoir en ces techniques ?

Devant la prolifération des algorithmes de classification et des questions méthodologiques et épistémologiques qu'elle soulève, les analyses comparatives sont devenues un enjeu de recherche important dans le domaine de la fouille de texte, de la fouille de données en général, mais également en sciences humaines et sociales. Les principales stratégies de comparaison sont brièvement présentées dans la section 2.

2. État de l'art sur les stratégies d'analyse comparative

L'analyse comparative des algorithmes de classification automatique s'est essentiellement fait selon trois stratégies : une comparaison basée sur des critères externes de validité, internes ou relatifs (Jain et al., 1999 : 268).

2.1. Comparaison basée sur des critères externes de validité

La première stratégie utilise une source externe, qui sert de benchmark pour la comparaison. La démarche dominante de cette stratégie consiste à comparer, à l'aide de taux de rappel, de taux de précision ou d'une combinaison des deux (la mesure-F par exemple), les résultats d'une classification automatique avec ceux d'une classification supervisée par des humains experts du domaine. Comparer deux algorithmes consiste alors à évaluer celui qui produit la partition la plus en adéquation avec le benchmark (Sebastiani, 2002).

2.2. Comparaison basée sur des critères internes de validité

La deuxième stratégie de comparaison consiste à vérifier, uniquement à partir d'information inhérente aux données textuelles étudiées, si celles-ci sont effectivement organisées selon une structure de classe. Cette stratégie s'appuie sur une définition particulière de la structure de classe, parfois appelée « classes naturelles » ou « hypersphériques ». Comparer différents algorithmes de classification automatique consiste dans ce cas à évaluer leur capacité à optimiser une fonction qui maximise la similitude intra-classe et la distance inter-classes (Manning et al., 2008 : 327; Weiss et al., 2005 : 123-124). On retrouve dans la littérature plusieurs dizaines de métriques qui ont été utilisées à cet égard : coefficient Silhouette, l'indice David-Bouldin, l'indice de Dunn, etc. Ces métriques sont basées sur des coefficients d'association, de densité, de diamètre des classes, d'écart aux centroïdes et d'autres (Vendramin et al., 2009).

2.3. Comparaison basée sur des critères relatifs

La troisième stratégie consiste à comparer, en fonction de leur similitude ou de leur variation, les partitions produites par les différents algorithmes de classification automatique évalués (Rand, 1971; Hubert and Arabie, 1985). Contrairement aux deux approches précédentes où l'évaluation est basée sur une norme externe – un benchmark pour la première et une fonction d'optimisation pour la seconde – cette troisième stratégie est une évaluation par triangulation ou analyse du consensus entre algorithmes.

Cette stratégie est particulièrement intéressante relativement aux questions de recherche soulevées précédemment. Le principe de la triangulation veut que, si on peut démontrer que différents algorithmes de classification génèrent sur un même ensemble d'objets des partitions similaires, il est alors raisonnable de conclure que les classes identifiées sont robustes et non un artefact des techniques utilisées (Milligan and Cooper, 1987 : 348). Les différentes techniques se corroborent les unes les autres. À l'inverse, s'il y a absence de similitude, la validité des résultats de la classification doit être questionnée.

Différentes métriques également sont utilisées pour mesurer la similitude entre deux partitions, certaines sont des mesures d'intersection d'autres de l'entropie (Vinh et al., 2009). Ce type de stratégie d'analyse comparative n'a été utilisé qu'en de très rares occasions pour la comparaison d'algorithmes de classifications appliqués au texte. Par contre, elle fait actuellement l'objet d'un intérêt particulier dans le domaine de la fouille de donnée et de la classification dite par « consensus » (Strehl and Ghosh, 2002; Fred and Jain, 2005). Les résultats issus de ces études sont cependant difficiles à généraliser au domaine de l'analyse de texte en sciences humaines et sociales. En effet, la classification est généralement appliquée sur des objets artificiels décrits par un très petit nombre de variables – moins d'une dizaine – alors que la classification textuelle s'applique au contraire sur des objets beaucoup plus complexes, qui ont souvent des centaines voire des milliers de variables.

3. Objectif et hypothèses de recherche

À notre connaissance, aucune étude systématique basée sur des critères relatifs de comparaison n'a été menée ni dans les domaines techniques de la fouille de texte, ni dans les sciences humaines et sociales. Notre objectif de recherche est d'entreprendre cette étude et d'évaluer l'importance du choix de l'algorithme dans une pratique d'analyse de texte assistée par ordinateur. Nous cherchons à mesurer la variation dans les résultats de classification en fonction de différents algorithmes. Deux hypothèses générales sont formulées sur la forme de cette variation.

3.1. Hypothèse sur la variation globale entre deux partitions

Sur un même ensemble d'objets, en l'occurrence un même corpus de textes, il y a toujours plusieurs partitions possibles et l'identification de l'une au détriment des autres dépend de plusieurs choix d'opérationnalisation dans la méthode : calibrage des paramètres de la segmentation du corpus, l'indexation des termes, la lemmatisation, l'application des antidictionnaires, la vectorisation, etc.

D'autre part, chaque algorithme de classification est basé sur un principe d'induction spécifique susceptible aussi de faire varier les résultats de la classification. Notre première hypothèse (H1) propose de vérifier si, toutes choses étant égales par ailleurs dans le prétraitement des données, les résultats de la classification varient significativement selon l'algorithme utilisé. Cette hypothèse a pour but d'évaluer de manière globale à quel point la classification textuelle est tributaire du choix d'un algorithme, et indirectement, la robustesse des classes identifiées par de telles techniques et le niveau de confiance que l'on peut avoir en elles.

3.2. Hypothèse sur les invariances locales entre deux partitions

Par ailleurs, les segments de texte composant un corpus possèdent leurs propres contraintes internes selon la distribution de leurs variables, lexèmes ou autres (Harris, 1991). Un algorithme de classification cherche à faire émerger de manière ascendante des classes en adéquation avec les contraintes internes de ces segments. Toutefois, ces contraintes varient pour chaque segment : certains sont regroupés de manière univoque, alors que d'autres, plus ambigus ou complexes, peuvent être classés de plusieurs manières équivoques.

Ainsi, il est possible que certaines contraintes dominantes dans un corpus puissent être identifiées peu importe le principe d'induction utilisé, autrement dit, que des patrons de classification de certains segments puissent être très similaires peu importe l'algorithme utilisé. Notre deuxième hypothèse (H2) concerne ces invariances locales. H2 propose que, toutes choses étant égales dans le prétraitement des données, la variation globale attendue dans H1 dissimule des invariances locales dans les patrons de classification de certains segments. Cette hypothèse a pour but d'évaluer de manière locale si la classification de certains segments de texte est plus robuste que d'autres et indépendante de l'algorithme utilisé. De plus, elle permet de moduler le niveau de confiance des chercheurs dans les techniques de classification automatique pour chaque segment de texte.

4. Expérimentation et premiers résultats

Nos premières expérimentations ont été menées avec 4 corpus textuels de test largement utilisés dans le domaine de la fouille de textes, soit la collection Reuters-21578, la collection Cranfield, la collection Medline et la collection Time.

Le prétraitement de ces quatre collections a consisté à ne conserver que le titre et le corps de texte de chaque segment et filtré toutes les métadonnées (étiquette, auteur, date, etc.). La ponctuation, les nombres, les singletons et les mots fonctionnels¹ ont été filtrés, les termes lemmatisés selon l'algorithme de Porter (1980). Les hapax et les termes distribués dans plus de 50% des segments ont également été filtrés. La pondération des variables est binaire (présence/absence) et la mesure de similarité entre segments est euclidienne (Tab. 1).

Corpus	N	$ X $	$ X' $	Poids	Similarité	K	Source
Reuters	3299	27182	6047	Binaire	Euclidienne	61	http://www.daviddlewis.com/resources/testcollections/reuters21578/
Cranfield	1398	10186	2629	Binaire	Euclidienne	37	http://ir.dcs.gla.ac.uk/resources/test_collections/cran/
Medline	1033	14052	4343	Binaire	Euclidienne	36	http://ir.dcs.gla.ac.uk/resources/test_collections/medl/
Time	423	36010	8389	Binaire	Euclidienne	25	http://ir.dcs.gla.ac.uk/resources/test_collections/time/

Tableau 1 : Récapitulatif des données utilisées pour les expérimentations, où N est le nombre de segments de texte, $|X|$ le nombre de variables (lexème), $|X'|$ le nombre de variables (lemme) après le prétraitement et K le nombre de classes générées

Nous présentons ici les premiers résultats de cette étude où sont mesurées les similitudes entre les partitions produites par 4 algorithmes : les réseaux à résonance adaptative (ART1; Carpenter et al., 1991) et les cartes auto-organisatrices (SOM; Kohonen, 2001), les K-Means (KM; MacQueen, 1967) et l'algorithme Expectation-Maximization (EM; Dempster et al., 1977). Chaque algorithme est appliqué aux 4 corpus retenus pour l'expérimentation. Pour la classification de chaque corpus, le nombre K de classes est fixé arbitrairement le plus près possible de \sqrt{N} , suivant l'heuristique de Bezdek et Pal (1998). Par exemple, à partir du corpus Time, ont été générés 4 partitions différentes de 23 classes chacune respectivement en utilisant l'algorithme ART1, SOM, KM et EM.

Les résultats d'une classification sur un corpus C de N segments de texte forme une partition P^u dont la structure peut être représentée par l'ensemble des relations $R = \{r_{ij}, r_{ij}, \dots, r_{mn}\}$ entre les M couples de segments, où $M = N \times N = |R|$. Une relation r_{ij} qui lie un couple de segments est positive (égale à 1) si les segments c_i et $c_j \in C$ sont classés ensemble par l'algorithme et négative (égale à 0) s'ils ne le sont pas.

La comparaison peut alors consister à mesurer l'intersection entre deux partitions générées à l'aide de deux algorithmes différents tel $P^u \cap P^v / P^u \cup P^v$. La notation conventionnelle du Tab. 2 représente les correspondances possibles entre deux partitions P^u et P^v :

P^u / P^v		P^v	
		Même classe	Classe différente
P^u	Même classe	a	b
	Classe différente	c	d

Tableau 2 : Table de contingence représentant les correspondances entre deux partitions

¹ L'antidictionnaire de mots fonctionnels utilisé est celui de Moreno (http://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/chercheurs/torres/logiciels/fonctionnels_eng.txt).

Dans laquelle : a est le nombre de couples de segments liés par une relation r_{ij} positive à la fois dans P^u et P^v ; b est le nombre de couples de segments liés par une relation r_{ij} positive dans P^u et négative dans P^v ; c le nombre de couples de segments liés par une relation r_{ij} négative dans P^u et positive dans P^v ; d le nombre de couples de segments liés par une relation r_{ij} négative à la fois dans P^u et P^v , et où $a+b+c+d = M = N \times N = |R|$.

La même démarche peut s'appliquer localement à chaque segment de texte. Le patron de classification S de chaque segment peut être représenté par un sous-ensemble de R tel $S \subset R$, $S \neq R$, $|S| = N$. Pour un segment donné, l'intersection $S^u \cap S^v$ correspond dans ce cas à la similitude entre les patrons de classification générés par différents algorithmes.

Pour nos expérimentations, nous avons retenu 3 différents calculs de l'intersection : via l'indice de Jaccard $J = a/a+b+c$, l'index de Rand $IR = a+d/a+b+c+d$ et le coefficient Kappa $K = 2(ad - bc) / ((a+b)(b+d)) + ((c+d)(a+c))$. Les valeurs de ces métriques se situent entre 0 et 1 pour Jaccard et Rand et entre -1 et 1 pour Kappa. Pour les trois métriques, une valeur ≤ 0 correspond à une absence de similitude entre les deux partitions comparées et une valeur de 1 à deux partitions identiques.

Ces 3 métriques sont basées sur une conceptualisation différente de l'intersection entre deux partitions. L'indice de Jaccard considère pertinente uniquement l'intersection positives de deux partitions, alors que l'index de Rand considère à la fois l'intersection positive et négative. Tandis que Jaccard est très restrictif, l'index de Rand est reconnu pour surestimer les similitudes entre deux partitions (Hubert and Arabie, 1985), car la probabilité que deux segments de texte ne soient pas classés de manière similaire par deux algorithmes est toujours plus grande qu'ils le soient. Pour éliminer ce biais, le coefficient Kappa quant à lui mesure l'intersection positive et négative observée entre deux partitions moins celle attendue par la chance. Ce qui fait d'ailleurs du coefficient Kappa, et de son équivalent dans l'index de Rand ajusté, l'une des métriques les plus utilisées à cet égard (Warrens, 2008).

4.1. Présentation des résultats

La présentation des résultats de l'expérimentation se divise en 2 sous-sections : les résultats de la première sous-section sont relatifs à l'hypothèse H1 sur la variation globale entre partitions générées par différents algorithmes; la deuxième sous-section présente les résultats relatifs à l'hypothèse H2 sur les invariances locales. Pour les N segments de chaque corpus a été appliquée une classification automatique de K classes à l'aide des 4 algorithmes retenus pour l'expérimentation. Donc, pour chaque corpus a été effectué 6 comparaisons : (i) ART1_SOM, (ii) ART1_KM, (iii) ART1_EM, (iv) SOM_KM, (v) SOM_EM et (vi) EM_KM.

4.1.1. Comparaisons globales des partitions

Dans Fig. 1 sont illustrés les résultats des analyses comparatives globales entre partitions, respectivement pour le corpus Reuters (a), le corpus Cranfield (b), le corpus Medline (c) et le corpus Time (d). Chaque histogramme illustre en fonction des 3 métriques Jaccard, Rand et Kappa, la valeur de la similitude obtenue pour les 6 comparaisons deux à deux entre algorithmes de classification automatique.

Pour le corpus Reuters, les similitudes entre les partitions produites par les algorithmes de classification automatique ART1, SOM, KM et EM sont de 0,06 à 0,14 ($\bar{x} = 0,10$) pour l'indice de Jaccard; de 0,85 à 0,91 ($\bar{x} = 0,88$) pour l'index de Rand; et de 0,07 à 0,22 ($\bar{x} = 0,12$) pour le coefficient Kappa. Pour le corpus Cranfield, les similitudes entre les partitions produites par

les 4 différents algorithmes de classification automatique sont de 0,05 à 0,11 ($\bar{x} = 0,08$) pour Jaccard; de 0,81 à 0,91 ($\bar{x} = 0,85$) pour Rand; et de 0,04 à 0,15 ($\bar{x} = 0,08$) pour Kappa. Pour le corpus Medline, les similitudes entre les partitions produites par les 4 différents algorithmes sont de 0,07 à 0,22 ($\bar{x} = 0,11$) pour Jaccard; de 0,71 à 0,89 ($\bar{x} = 0,80$) pour Rand; et de 0,05 à 0,24 ($\bar{x} = 0,11$) pour Kappa. Pour le corpus Time, les similitudes entre les partitions produites par les algorithmes de classification automatique ART1, SOM, KM et EM sont de 0,06 à 0,39 ($\bar{x} = 0,12$) pour Jaccard; de 0,44 à 0,90 ($\bar{x} = 0,63$) pour Rand; et de 0,02 à 0,26 ($\bar{x} = 0,08$) pour Kappa.

Pour les 4 corpus, la similitude moyenne mesurée avec Kappa entre les partitions produites par ART1 et SOM est de 0,08; entre ART1 et KM de 0,05; entre ART1 et EM de 0,06; entre SOM et KM de 0,12; entre SOM et EM de 0,10 et entre EM et KM de 0,19. En moyenne, l'intersection entre les partitions de ART1 et celles générées par les autres algorithmes est de 0,06 toujours selon Kappa; l'intersection moyenne des partitions de SOM avec celles des autres algorithmes est de 0,10; l'intersection moyenne de KM est de 0,12 et celle de EM est de 0,11.

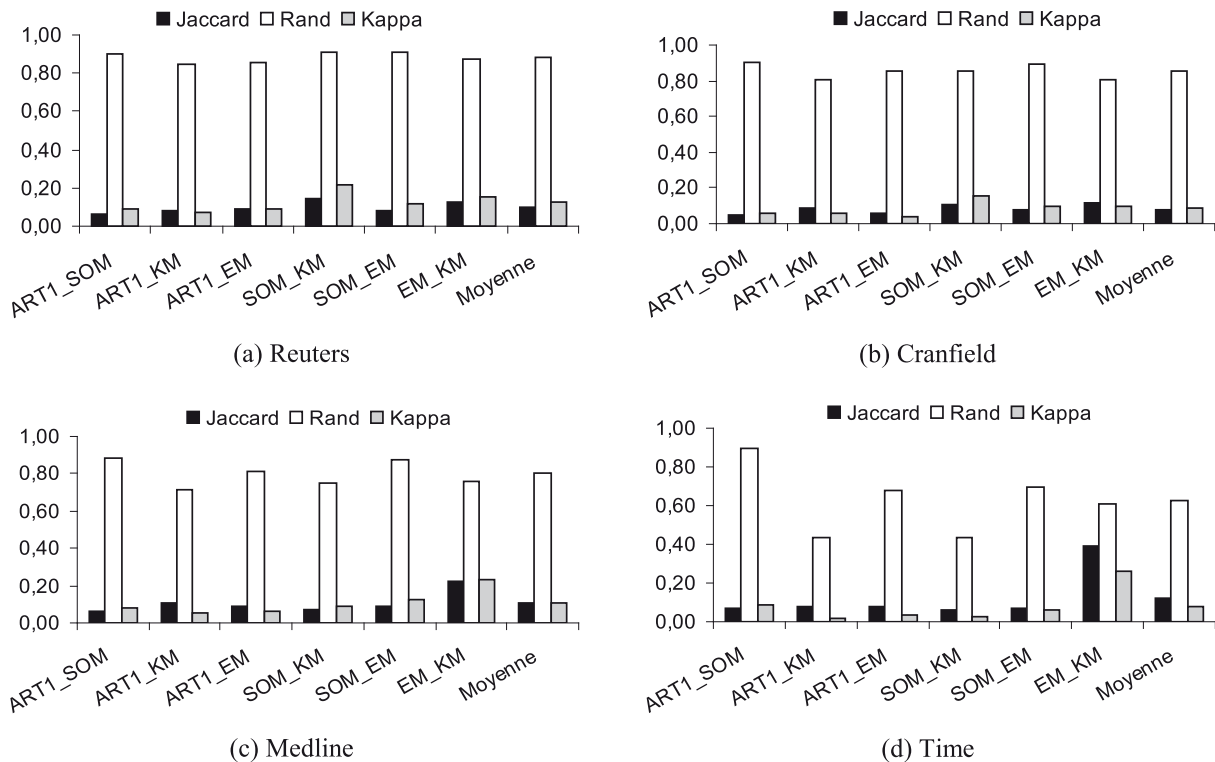


Figure 1 : Histogrammes illustrant, pour les 4 corpus, les similitudes globales pour les comparaisons des partitions générées par les algorithmes de classification automatique ART1, SOM, KM et EM

4.1.2. Comparaisons locales des patrons de classification de chaque segment

Dans Fig. 2 sont illustrés les résultats des analyses comparatives locales entre patrons de classification pour chaque segment des corpus Reuters (a), Cranfield (b), Medline (c) et Time (d). La similitude est mesurée avec le coefficient Kappa et les 4 algorithmes de classification automatique ART1, SOM, KM et EM sont comparés. Chaque figure contient un diagramme en boîte qui illustre la variation de la valeur du Kappa pour les comparaisons deux à deux de tous les patrons de classification de chaque segment.

Pour le corpus Reuters, l'étendue de la variation du Kappa pour tous les patrons de classification de chaque segment de chaque algorithme est de -0,21 à 1,00 avec $\bar{x} = 0,15$ et $\sigma = 0,18$. Selon les algorithmes comparés, le dernier décile contient des valeurs de Kappa en moyenne supérieures à 0,39. Pour le corpus Cranfield, l'étendue de la variation du Kappa pour tous les patrons de classification de chaque segment de chaque algorithme est de -0,18 à 0,67 avec $\bar{x} = 0,10$ et $\sigma = 0,12$. Selon les algorithmes comparés, le dernier décile contient des valeurs de Kappa en moyenne supérieures à 0,24. Pour le corpus Medline, l'étendue de la variation du Kappa pour tous les patrons de classification de chaque segment de chaque algorithme est de -0,25 à 0,67 avec $\bar{x} = 0,12$ et $\sigma = 0,16$. Selon les algorithmes comparés, le dernier décile contient des valeurs de Kappa en moyenne supérieures à 0,38. Pour le corpus Time, l'étendue de la variation du Kappa pour tous les patrons de classification de chaque segment de chaque algorithme est de -0,15 à 1,00 avec $\bar{x} = 0,09$ et $\sigma = 0,14$. Selon les algorithmes comparés, le dernier décile contient des valeurs de Kappa en moyenne supérieures à 0,20.

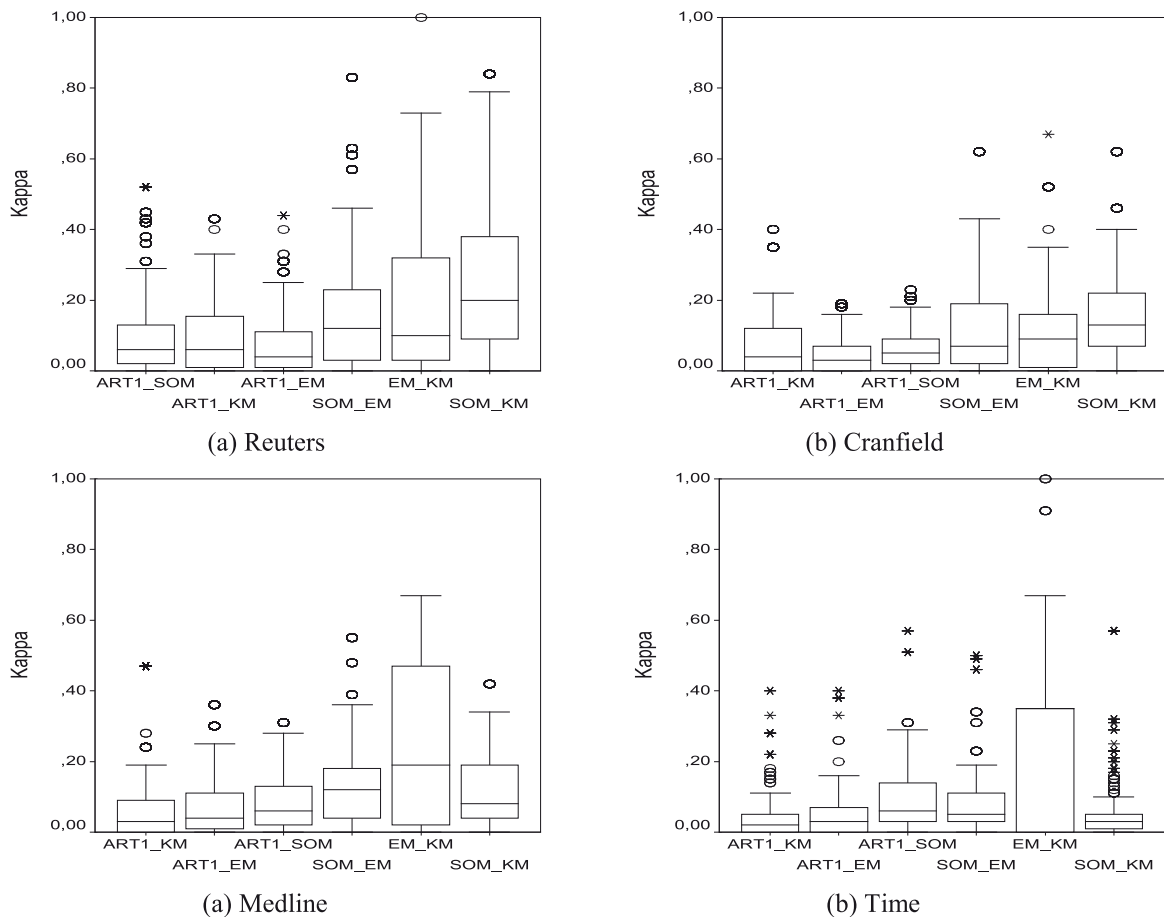


Figure 2 : Diagrammes en boîte illustrant la variation, pour les 4 corpus, de la similitude entre patrons de classification générés par les algorithmes de classification automatique ART1, SOM, KM et EM

5. Discussion

Ces premiers résultats nous permettent de revenir commenter nos 2 hypothèses de recherche. La classification automatique est souvent présentée comme une technique qui, parce qu'ascendante et faiblement contraignante, permet de faire émerger une structure de classe inhérente aux données textuelles analysées sans leur imposer de manière externe des formes prédéfinies.

Or, conformément à l'hypothèse H1, nos résultats de recherches montrent le contraire. L'hypothèse H1 proposait que, toutes choses étant égales par ailleurs dans le prétraitement des données textuelles, le choix de l'algorithme était déterminant sur les résultats de la classification automatique. Les résultats présentés dans la section 4.3.1 corroborent fortement cette hypothèse. On y observe que les similitudes entre différentes partitions obtenues avec différents algorithmes sont en moyenne seulement de 0,10 selon l'indice de Jaccard; 0,79 selon l'index de Rand; et de 0,12 selon le coefficient de Kappa ². Par ailleurs, on remarque également que les partitions générées par l'algorithme ART1 sont plus idiosyncrasiques que les autres, alors que celles produites par KM sont celles qui croisent le plus les autres. Les plus grandes différences sont entre les partitions de ART1 et KM, et les plus grandes ressemblances sont entre EM et KM.

D'autres parts, l'hypothèse H2 nous avait permis de formuler l'idée selon laquelle certains segments de texte composant un corpus sont classés de manière plus univoque que d'autres, indépendamment de l'algorithme utilisé. Une analyse par triangulation telle que nous l'avons faite devait montrer que certains patrons de classification de certains segments sont très similaires d'un algorithme à l'autre. Les résultats présentés dans la section 4.3.2 corroborent partiellement cette hypothèse. On y observe que pour chaque segment, la similitude entre patrons de classification de différents algorithmes varie beaucoup, soit de -0,25 à 1,00. On remarque que, peu importe le corpus et l'algorithme, un petit noyau de segments est classé de manière relativement analogue : le dernier décile de la distribution des valeurs du Kappa varie en moyenne entre 0,30 et 0,54. Ceci signifie que sur l'ensemble d'un corpus de texte soumis à la classification, 10% des segments vont avoir des patrons de classification beaucoup plus robuste que les autres, c'est-à-dire vont être invariablement regroupés avec les mêmes segments sans égard à l'algorithme utilisé.

6. Conclusion : le poids des choix d'opérationnalisation

Les chercheurs des sciences humaines et sociales qui pratiquent l'analyse textuelle à l'aide de technique de classification automatique utilisent une méthode complexe qui implique plusieurs choix d'opérationnalisation qui détermineront leur parcours d'exploration des données et leurs résultats d'analyse. L'un des choix d'opérationnalisation est l'algorithme de classification. L'objectif de cette étude était d'évaluer à quel point les résultats de la classification en sont tributaires. Nos expérimentations ont été menées sur 4 corpus et 4 algorithmes différents. Nos résultats de recherche nous amènent à la conclusion que ce choix d'opérationnalisation est déterminant. Seulement un consensus relatif d'environ 10% entre les partitions produites par différents algorithmes est observé.

Les conclusions de cette recherche doivent néanmoins être généralisées avec prudence. Les résultats sont circonscrits aux 4 corpus étudiés, qui ont peut-être, des caractéristiques spécifiques qui les rendent difficiles à classer, ou du moins, dont la complexité est telle qu'il est illusoire d'identifier une structure de classe univoque. De plus, des expérimentations supplémentaires

² Comme on pouvait s'y attendre, l'index de Rand surestime largement les similitudes entre partitions. Les valeurs élevées de l'index de Rand pourraient laisser supposer que l'intersection entre les différentes partitions ne porte pas tant sur leurs intersections positives que sur leurs intersections négatives. Toutefois, en contrôlant la part de ces similitudes due à la chance, les valeurs du coefficient Kappa rejoignent celles de l'indice de Jaccard. Ceci signifie que les intersections négatives entre deux partitions sont de très mauvais indicateurs de leurs similitudes.

doivent être menées en faisant varier le nombre K de classes selon différentes heuristiques. Différents prétraitements des données, en particulier dans la sélection des variables (lexème, lemme, n-gramme, etc.), peuvent aussi orienter les résultats, quoique, le ratio nombre de variables / nombre de segments ne semble pas corrélé aux observations effectuées.

Par ailleurs, cette étude ne dit rien sur quel algorithme est le « meilleur » ; elle tient pour acquis qu'ils ont déjà amplement fait leur preuve, certains depuis plus de 40 ans. Les conséquences méthodologiques et épistémologiques de cette étude ne sont pas moins importantes. Elles indiquent que, appliqué sur des données complexes comme le texte, un algorithme de classification automatique individuellement ne peut épuiser l'espace des classifications possibles. Elles indiquent également que la confiance des chercheurs en ces techniques ne peut être tenue pour acquise. Elle peut cependant être accompagnée par différents mécanismes de contrôle comme l'évaluation par triangulation que nous avons menée.

Ce constat est d'autant plus pertinent pour les sciences humaines et sociales où la classification automatique est souvent appliquée sur des corpus sémantiquement complexes. Il n'est pas contradictoire que deux partitions radicalement différentes soient malgré tout valides. En effet, rares sont les problématiques de recherche où un sociologue ou un philosophe par exemple, sera amené à faire de la classification automatique sur un corpus constitué à la fois de recettes de cuisine française, d'articles sur la crise financière et de l'œuvre de Darwin. Si nous avions fait une telle expérimentation, probablement que nous aurions eu des consensus plus fort pour indiquer qu'une recette de tartare, la baisse des taux d'intérêt et le concept d'évolution ne sont pas équivalents. Cependant, les problématiques de recherche en sciences humaines et sociales sont souvent toutes autres et les classes d'équivalence recherchées sont sémantiquement subtiles et ambiguës, ce qui, d'ailleurs, en fait un objet de découverte.

Références

- Berkhin P. (2006). Survey of clustering data mining techniques. In Kogan, J., Nicholas, C. and Teboulle, M., editors, *Grouping Multidimensional Data*, Berlin-Heidelberg: Springer, pp. 25-71.
- Bezdek J.C. and Pal N.R. (1998). Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics—part b: Cybernetics*, 28, 3 : 301-315.
- Carpenter G., Grossberg A.S. and Rosen D.B. (1991). ART 2-A: An adaptive resonance algorithm for rapid category learning and recognition. *Neural Networks*, 4,4 : 493-504.
- Demazière D., Brossaud C., Trabal P. and Van Meter K. (editors) (2006). *Analyses textuelles en sociologie - Logiciels, méthodes, usages*. Renne : PUR.
- Dempster A.P., Laird N.M. and Rubin D.B. (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society*, B, 39 : 1-38.
- Diesner J. and Carley K.M. (2005). Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis. In Narayanan, V.K. and Armstrong, D.J., editors, *Causal Mapping for Information Systems and Technology Research: Approaches, Advances, and Illustrations*, Harrisburg, PA: Idea Group Publishing, pp. 81-108.
- Estivill-Castro V. (2002). Why so many clustering algorithms: a position paper. *ACM SIGKDD Explorations Newsletter*, 4, 1 : 65-75.
- Forest D. and Meunier J.-G. (2004). Classification et catégorisation automatiques : application à l'analyse thématique des données textuelles. In Purnelle, G., Fairon, C. and Dister, A., editors, *JADT2004*, 10-12 mars, Presses universitaires de Louvain, pp. 434-444.

- Fred A.L.N and Jain A.K. (2005). Combining multiple clusterings using evidence accumulation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27, 6 : 835-850.
- Gagnon D. (2004). NUMEXO et l'analyse par attracteurs par classes des entrées de l'ECHO (Encyclopédie Culturelle hypermédia de l'Océanie). *Lexicometrica, L'analyse de données textuelles : De l'enquête aux corpus littéraires*, Numéro spécial.
- Harris Z.S. (1991). *A Theory of Language and Information: A Mathematical Approach*. Oxford : Clarendon Press.
- Hockey S. (2001). *Electronic Texts in the Humanities: Principles and Practice*. Oxford : Oxford University Press.
- Hubert L. and Arabie P. (1985). Comparing partitions. *Journal of Classification*, 2 : 193-218.
- Jain A.K., Murty M.N. and Flynn P.J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31, 3 : 264-323.
- Kalampalikis N. and Moscovici S. (2005). Une approche pragmatique de l'analyse Alceste. *Les Cahiers Internationaux de Psychologie Sociale*, 66 : 15-24.
- Kohonen T. (2001). *Self-Organizing Maps*. Berlin: Springer.
- Lalhou S. (2003). L'exploration des représentations sociales à partir des dictionnaires. In Abric, J.C., editor, *Méthodes d'étude des représentations sociales*, Ramonville Saint-Agne : Eres, pp. 37-58.
- Lebart L. (2004). Validité des visualisations de données textuelles. In Purnelle, G., Fairon, C., and Dister, A., editors, *JADT2004*, 10-12 mars, Presses universitaires de Louvain, pp. 708-715.
- MacQueen J.B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings Berkeley Symposium on Mathematics, Statistics and Probability*, University of California Press, pp. 281-297.
- Manning C.D., Raghavan P. and Schütze H. (2008). *Introduction to Information Retrieval*. Cambridge : Cambridge University Press.
- Meunier J.-G. and Forest D. (2009). L'analyse conceptuelle assistée par ordinateur: premières expériences. In Le Priol, F., Djioua, B. and Desclés, J.-P., editors, *L'annotation*, Paris : Hermès, pp. 211-230.
- Meunier J.-G., Forest D. and Biskri I. (2005). Classification and categorization in computer assisted reading and analysis of texts. In Cohen, H. and Lefebvre, C., editors, *Handbook of Categorization in Cognitive Science*, Amsterdam : Elsevier, pp. 955-978.
- Milligan G.W. and Cooper M.C. (1987). Methodology review: clustering methods. *Applied Psychological Measurement*, 11, 4 : 329-354.
- Porter M.F. (1980). An algorithm for suffix stripping. *Program*, 14 : 130-137.
- Rand W.M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66, 336 : 846-850.
- Rastier F. (2001). *Arts et sciences du texte*. Paris : PUF.
- Reinert M. (1993). Les mondes lexicaux et leur logique. *Langage et Société*, 66 : 5-39.
- Salton G. (1989). *Automatic Text Processing*. Reading (MA) : Addison-Wesley.
- Sebastiani F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34, 1 : 1-47.
- Strehl A. and Ghosh J. (2002). Cluster ensembles – A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3 : 583-617.
- Vendramin L., Campello R.J.G.B. and Hruschka E.R. (2009). On the comparison of relative clustering

- validity criteria. In *Proceedings of the Ninth SIAM International Conference on Data Mining*, pp. 733-744.
- Vinh N.X., Epps J. and Bailey J. (2009). Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th International Conference on Machine Learning*, Montréal, Canada, pp. 1073-1080.
- Warrens M.J. (2008). On the equivalence of Cohen's Kappa and the Hubert-Arabie adjusted rand index. *Journal of Classification*, 25 : 177-183.
- Weiss S.M., Indurkha N., Zhang T. and Damereau F.J. (2005). *Text Mining. Predictive Methods for Analyzing Unstructured Information*. New York : Springer.