Une nouvelle approche pour la sélection de variables basée sur une métrique d'estimation de la qualité

Jean-Charles Lamirel*, Pascal Cuxac**, Kafil Hajlaoui**

*SYNALP Team-LORIA, INRIA Nancy-Grand Est, Vandoeuvre-lès-Nancy, France.
jean-charles.lamirel@loria.fr,
http://www.loria.fr
**CNRS-Inist, Vandoeuvre-lès-Nancy, France.
pascal.cuxac@inist.fr
http://www.inist.fr

Résumé. La maximisation d'étiquetage (F-max) est une métrique non biaisée d'estimation de la qualité d'une classification non supervisée (clustering) qui favorise les clusters ayant une valeur maximale de F-mesure d'étiquetage. Dans cet article, nous montrons qu'une adaptation de cette métrique dans le cadre de la classification supervisée permet de réaliser une sélection de variables et de calculer pour chacune d'elles une fonction de contraste. La méthode est expérimentée sur différents types de données textuelles. Dans ce contexte, nous montrons que cette technique améliore les performances des méthodes de classification de façon très significative par rapport à l'état de l'art des techniques de sélection de variables, notamment dans le cas de la classification de données textuelles déséquilibrées, fortement multidimensionnelles et bruitées.

1 Introduction

Depuis les années 1990, les progrès de l'informatique et des capacités de stockage permettent la manipulation de très gros volumes de données : il n'est pas rare d'avoir des espaces de description de plusieurs milliers, voire de dizaines de milliers de variables. On pourrait penser que les algorithmes de classification sont plus efficaces avec un grand nombre de variables, mais la situation n'est pas aussi simple que cela. Le premier problème qui se pose est l'augmentation du temps de calcul. En outre, le fait qu'un nombre important de variables soit redondant ou non pertinent pour la tâche de classification perturbe considérablement le fonctionnement des classifieurs. De plus, la plupart des algorithmes d'apprentissage exploitent des probabilités dont les distributions peuvent être difficiles à estimer en présence d'un très grand nombre de variables. L'intégration d'un processus de sélection de variables dans le cadre de la classification des données de grande dimension devient donc un enjeu central. Dans la littérature, trois types d'approches pour la sélection de variables sont principalement proposés : les approches directement intégrées aux méthodes de classification, dites «embedded», les méthodes basées sur des techniques d'optimisation, dites «wrapper», et finalement, les approches de filtrage. Des états de l'art exhaustifs ont été réalisés par de nombreux auteurs, comme Ladha

et al. (Ladha et Deepa, 2011), (Bolón-Canedo et al., 2012) ou (Guyon et Elisseeff, 2003). Nous ne faisons donc ci-après qu'un rapide tour d'horizon des approches existantes.
Les approches « embedded » intègrent la sélection des variables dans le processus d'apprentissage (Breiman et al., 1984). Les méthodes les plus populaires de cette catégorie sont les méthodes basées sur les SVM et les méthodes fondées sur les réseaux de neurones. A titre d'exemple, SVM-EFR (Recursive Feature Elimination for Support Vector Machines) (Guyon et al., 2002) est un processus intégré qui effectue la sélection des variables de façon itérative en utilisant un classificateur SVM et en supprimant les variables les plus éloignées de la frontière de décision.
De leur côté, les méthodes « wrappers » utilisent un critère de performance pour la recherche d'un sous-ensemble de prédicteurs pertinents (Kohavi et John, 1997). Le plus souvent, c'est le taux d'erreur (mais cela peut être un coût de prédiction ou l'aire sous la courbe ROC). A titre d'exemple, la méthode WrapperSubsetEval commence avec un ensemble vide de variables et se poursuit jusqu'à ce que l'ajout de nouvelles variables n'améliore plus les performances, en exploitant la validation croisée pour estimer la précision de l'apprentissage pour un ensemble donné de variables (Witten et Frank, 2005). Les comparaisons entre méthodes, comme celle de Forman (Forman, 2003), mettent clairement en évidence que, sans tenir compte de leur efficacité, l'un des principaux inconvénients de ces deux catégories de méthodes est qu'elles sont très gourmandes en temps de calcul. Cela proscrit leur utilisation dans le cas de données fortement multidimensionnelles. Dans ce contexte, une alternative possible est alors d'exploiter les méthodes de filtrage.
Les approches par filtrage sont des méthodes de sélection qui sont utilisées en amont et indépendamment de l'algorithme d'apprentissage. Basées sur des tests statistiques, elles sont plus légères en termes de temps de calcul que les autres approches.
La méthode du chi-carré exploite un test statistique courant qui mesure l'écart à une distribution attendue en supposant que les variables sont indépendantes des étiquettes de classe (Ladha et Deepa, 2011). Le gain d'information est également l'une des méthodes les plus courantes de l'évaluation de variables. Ce filtre univarié fournit une classification ordonnée de toutes les variables. Dans cette approche, les variables retenues sont celles qui obtiennent une valeur positive du gain d'information (Hall et Smith, 1999).
Dans la méthode MIFS (Mutual Information Feature Selection), une variable est ajoutée à un sous-ensemble de variables déjà sélectionnées si son lien avec la classe cible surpasse la connexion moyenne avec les prédicteurs déjà sélectionnés. La méthode prend en compte à la fois la pertinence et la redondance (Hall et Smith, 1999).
La méthode CBF (Consistency-based Filter) évalue la pertinence d'un sous-ensemble de variables par le niveau de cohérence des classes lorsque les échantillons d'apprentissage sont projetés sur ce sous-ensemble (Dash et Liu, 2003).
La méthode MODTREE est un procédé de filtrage qui repose sur le principe du calcul de la corrélation par paire. Elle fonctionne dans l'espace des paires d'individus décrits par des indicateurs de co-étiquetage attachés à chaque variable d'origine. Pour cela, un coefficient de corrélation par paire, qui représente la corrélation linéaire entre deux éléments, est utilisé. Le calcul des coefficients de corrélation partiels permet alors d'effectuer une sélection de variables pas à pas (Lallich et Rakotomalala, 2000).
L'hypothèse de base de la méthode Relief, qui tire son inspiration du principe des plus proches voisins, est de considérer une variable pertinente si elle discrimine bien un objet dans la classe

positive par rapport à son voisin le plus proche dans la classe négative. Le score des variables est cumulatif et calculé grâce à un tirage aléatoire de données-échantillons. ReliefF, une extension de Relief, ajoute la capacité de résoudre les problèmes multi-classes. Cette variante est aussi plus robuste et capable de traiter des données incomplètes et bruitées (Konokenko, 1995). ReliefF est considérée comme l'une des méthodes de sélection à base de filtres les plus efficaces.
Comme tout test statistique, les approches par filtrage sont connues pour avoir un comportement erratique dans le cas de variables de très faibles fréquences ; ce qui représente une situation habituelle dans la classification de texte (Ladha et Deepa, 2011). Nous montrons également dans cet article que, malgré leur diversité, toutes les approches de filtrages existantes s'avérent inopérantes, voir néfastes, dans le cas de données très déséquilibrées, fortement multidimensionnelles et bruitées, avec un degré de similitude élevé entre classes. Nous proposons comme alternative une nouvelle méthode de sélection de variables et de contraste basée sur la métrique de maximisation d'étiquetage, récemment développée, et nous comparons ses performances avec des techniques classiques dans le contexte d'aide à la validation des brevets. Nous étendons ensuite la portée de notre étude à des données textuelles de référence habituellement utilisées. La suite du document est structurée comme suit. La section 2 présente notre nouvelle approche de sélection de variables. La section 3 détaille les données utilisées. La section 4 compare les résultats de la classification avec et sans l'utilisation de l'approche proposée sur les différents corpus de données. La section 5 présente nos conclusions et perspectives.

2 Maximisation d'étiquetage pour la sélection de variables

La maximisation d'étiquetage (F-max) est une métrique non biaisée d'estimation de la qualité d'une classification non supervisée qui exploite les propriétés des données associées à chaque cluster sans examen préalable des profils de clusters (Lamirel et al., 2004). Son principal avantage est d'être tout à fait indépendante des méthodes de classification et de leur mode opératoire. Lorsqu'elle est utilisée après l'apprentissage, elle peut être exploitée pour établir des indices globaux de qualité de clustering (Lamirel et al., 2010) ou pour l'étiquetage de clusters (Lamirel et Ta, 2008). Considérons un ensemble de clusters C résultant d'une méthode de clustering appliquée sur un ensemble de données D représentées par un ensemble de variables F. La métrique de maximisation d'étiquetage favorise les clusters avec une valeur maximale de F-mesure d'étiquetage. La F-mesure d'étiquetage $FF_c(f)$ d'une variable f associée à un cluster c est définie comme la moyenne harmonique du rappel d'étiquetage $FR_c(f)$ et de la précision d'étiquetage $FP_c(f)$, eux-mêmes définis comme suit :

$$FR_c(f) = \frac{\Sigma_{d \in c} W_d^f}{\Sigma_{c' \in C} \Sigma_{d \in c'} W_d^f} \quad FP_c(f) = \frac{\Sigma_{d \in c} W_d^f}{\Sigma_{f' \in F_c, d \in c} W_d^{f'}} \tag{1}$$

avec

$$FF_c(f) = 2 \left(\frac{FR_c(f) \times FP_c(f)}{FR_c(f) + FP_c(f)} \right) \tag{2}$$

où W_d^f représente le poids de la variable f pour la donnée d et F_c représente l'ensemble des variables représentées dans les données associées au cluster c.

Tenant compte de la définition de base de la métrique de maximisation d'étiquetage, son exploitation pour la tâche de sélection de variables dans le contexte de l'apprentissage supervisé devient un processus simple, dès lors que cette métrique générique peut s'appliquer sur des données associées à une classe aussi bien qu'à celles qui sont associées à un cluster. Le processus de sélection peut donc être défini comme un processus non paramétré basé sur les classes dans lequel une variable de classe est caractérisée en utilisant à la fois sa capacité à discriminer une classe donnée ($FP_c(f)$ index) et sa capacité à représenter fidèlement les données de la classe ($FR_c(f)$ index). L'ensemble Sc des variables qui sont caractéristiques d'une classe donnée c, appartenant à un ensemble de classes C, se traduit par :

$$S_c = \left\{ f \in F_c \mid FF_c(f) > \overline{FF}(f) \text{ et } FF_c(f) > \overline{FF}_D \right\} \text{ où} \tag{3}$$

$$\overline{FF}(f) = \Sigma_{c' \in C} \frac{FF_{c'}(f)}{|C_{/f}|} \text{ et } \overline{FF}_D = \Sigma_{f' \in F} \frac{\overline{FF}(f')}{|F|} \tag{4}$$

où $C_{/f}$ représente le sous-ensemble de C dans lequel la variable f est représentée. Enfin, l'ensemble de toutes les variables S_C sélectionnées est le sous-ensemble de F défini comme :

$$S_C = \cup_{c \in C} S_C. \tag{5}$$

En d'autres termes, les variables qui sont jugées pertinentes pour une classe donnée sont les variables dont les représentations sont meilleures dans cette classe que leurs représentations moyennes dans toutes les classes, et meilleures que la représentation moyenne de toutes les variables, en termes de F-mesure d'étiquetage.
Dans le cadre spécifique du processus de maximisation d'étiquetage, une étape d'amélioration par contraste peut être exploitée en complément de la première étape de sélection. Le rôle de cette étape est d'adapter la description de chaque donnée aux caractéristiques spécifiques de leurs classes associées. Cela consiste à modifier le schéma de pondération des données pour chaque classe en prenant en considération le gain d'information fourni par la F-mesure d'étiquetage des variables, localement à cette classe. Le gain d'information est proportionnel au rapport entre la valeur de la F-mesure d'une variable dans la classe $FF_c(f)$ et la valeur moyenne de la F-mesure de cette variable sur l'ensemble de la partition $FF(f)$. Pour une donnée et pour une variable décrivant cette donnée, le gain résultant agit comme un facteur de contraste modulant le poids existant de cette variable dans le profil de la donnée, quel qu'il soit établi auparavant. Pour une variable f appartenant à l'ensemble S_c des variables sélectionnées d'une classe c, le gain $G_c(f)$ est exprimé comme suit :

$$G_c(f) = (FF_c(f)/\overline{FF}(f))^k \tag{6}$$

où k est un facteur d'amplification qui peut être optimisé en fonction de la précision obtenue.

Les variables actives d'une classe sont celles pour lesquelles le gain d'information est supérieur à 1 dans celles-ci. Etant donné que la méthode proposée est une méthode de sélection et de contraste basée sur les classes, le nombre moyen de variables actives par classe est donc comparable au nombre total de variables sélectionnées dans le cas des méthodes de sélection usuelles.

3 Données expérimentales

Un des buts poursuivi par le projet QUAERO est celui d'exploiter les informations bibliographiques pour aider des experts à juger de l'antériorité des brevets. Il s'agit donc, dans un premier temps, de prouver qu'il est possible d'associer ces informations de manière pertinente aux classes de brevets, autrement dit de les classifier correctement dans ces classes. Nos données source expérimentales principales contiennent 6387 brevets au format XML du domaine pharmacologique, regroupés en 15 sous-classes de la classe A61K (préparation médicale). Les citations bibliographiques dans les brevets sont extraites de la base de données Medline [1]. 25887 citations ont été extraites à partir de ces 6387 brevets. L'interrogation de la base de données Medline avec les citations extraites permet de récupérer les notices bibliographiques de 7501 articles. Chaque notice est ensuite marquée par le premier code de classement du brevet citant (Hajlaoui et al., 2012). Le résumé de chaque notice est traité et transformé en un sac de mots (Salton, 1971) en utilisant l'outil TreeTagger (Schmid, 1994). Pour réduire le bruit généré par cet outil, un seuil de fréquence de 45 (soit le seuil moyen de 3/classe) est appliqué sur les descripteurs extraits. Il en résulte un espace de description seuillé de dimension 1804. Une dernière étape de pondération TF-IDF (Salton, 1971) est appliquée. La série de notices étiquetées et ainsi pré-traitées représente le corpus final sur lequel l'apprentissage est effectué. Ce dernier corpus est fortement déséquilibré, la plus petite classe contenant 22 articles (classe A61K41) et la plus grande en contenant 2500 (classe A61K31). La similarité inter-classes calculée en utilisant une corrélation cosinus indique que plus de 70% des couples de classes ont une similitude comprise entre 0,5 et 0,9. Ainsi, la capacité d'un modèle de classification à détecter précisément la bonne classe est fortement réduite. Une solution habituellement utilisée pour faire face à un déséquilibre dans des données des classes est un sous-échantillonnage des grosses classes (Good, 2006) et/ou un sur-échantillonnage des petites classes (Chawla et al., 2002). Toutefois, le ré-échantillonnage, qui introduit de la redondance dans les données, n'améliore pas les performances avec cet ensemble de données, comme cela a été montré par Hajlaoui et al. (2012). Nous proposons donc ci-après, une solution alternative qui est d'élaguer les variables jugées non pertinentes et de contraster celles jugées fiables.

A titre complémentaire, 3 ensembles de données textuelles de référence sont également utilisés dans nos expériences :

- Les corpus R8 et R52 sont des corpus obtenus par Cardoso Cachopo [2] à partir des ensembles de données R10 et R90 issus de la collection Reuters 21578 [3]. Le but de ces adaptations est de ne retenir que les données ayant une seule étiquette. Considérant uniquement les documents monothématiques et les classes qui ont encore au moins un exemple d'apprentissage et un exemple de test, R8 est une réduction à 8 classes du corpus R10 (10 classes plus fréquentes) et R52 est une réduction à 52 classes du corpus R90 (90 classes).
- Le corpus Amazon^{tm} (AMZ) est un ensemble de données UCI (Bache et Lichman, 2013) dérivé des avis de clients du site web Amazon et exploitable pour l'identification des auteurs. Pour évaluer la robustesse des algorithmes de classification à un grand nombre de classes cibles, 50 des utilisateurs les plus actifs qui ont fréquemment postés des com-

1. http://www.ncbi.nlm.nih.gov/pubmed/
2. http://web.ist.utl.pt/~acardoso/datasets/
3. http://www.research.att.com/~lewis/reuters21578.html

mentaires dans ces newsgroups sont identifiés. Le nombre de messages collectés pour chacun d'entre eux est de 30. Chaque message comprend le style linguistique des auteurs tels que l'utilisation de chiffres, la ponctuation, les mots et les phrases fréquentes.

4 Expériences et résultats

4.1 Expériences

Pour effectuer nos expériences nous prenons d'abord en considération différents algorithmes de classification qui sont mis en oeuvre dans la boite à outils Weka [4] : arbres de décision (J48) (Quinlan, 1993), forêts aléatoires (RF) (Breiman, 2001), k-plus-proches-voisins (KNN) (Aha et al., 1991), des algorithmes bayésiens usuels, à savoir, bayésien naïf multinomial (MNB) et réseau bayésien (BN), et enfin, l'algorithme SMO-SVM (SMO) (Platt, 1999). Les paramètres par défaut sont utilisés lors de l'exécution de ces algorithmes, à l'exception de KNN pour lequel le nombre de voisins est optimisé sur la base de la précision résultante. Nous mettons ensuite plus particulièrement l'accent sur les tests d'efficacité des approches de sélection de variables, y compris notre nouvelle proposition (FMC). Nous incluons dans notre test un panel d'approches de filtrage qui sont applicables avec des données de grande dimension, en exploitant une nouvelle fois la plateforme Weka. L'ensemble des méthodes testées comprend : chi-carré (CHI), gain d'information (GI), CBF, incertitude symétrique (IS) (Yu et Liu, 2003), ReliefF (RLF), Analyse en Composantes Principales (PCA) (Pearson, 1901). Les paramètres par défaut sont utilisés pour la plupart de ces méthodes, sauf pour PCA pour lequel le pourcentage de variance expliquée est accordé en fonction de la précision obtenue. Dans un premier temps nous expérimentons les méthodes séparément. Dans une deuxième phase, nous combinons la sélection des variables fournies par les différentes méthodes avec la méthode de contraste que nous avons proposée (eq. 6). Une validation croisée en 10 feuillets (10-fold cross-validation) est utilisée dans l'ensemble de nos expériences.

4.2 Résultats

Les différents résultats sont présentés dans les tableaux 1 à 8. Ils se basent sur les mesures de performance standard (taux de vrai positif (TP) ou Rappel (R), taux de faux positifs (FP), Précision (P), F-mesure (F) et ROC) pondérées par la taille des classes, puis moyennés sur toutes les classes. Pour chaque table et chaque combinaison de méthodes de sélection et de classification, un indicateur de gain/perte de performance (TP Incr) est calculé en utilisant le taux TP de SMO sur les données originales comme référence. Enfin, comme les résultats s'avèrent identiques pour chi-carré, gain d'information et incertitude symétrique, ils ne figurent qu'une seule fois dans les tableaux comme résultats de type chi-carré (et sont notés CHI+). Pour notre collection principale de brevets, le tableau 1 met en évidence que les performances de toutes les méthodes de classification sont faibles sur l'ensemble de données considéré si aucun processus de sélection de variables n'est exécuté. Il confirme également dans ce contexte la supériorité des méthodes SMO, KNN et bayésiennes sur les deux autres méthodes basées sur les arbres de décision. En outre, SMO fournit la meilleure performance globale en termes de discrimination comme le montre sa valeur de ROC la plus élevée. Toutefois, la méthode

4. http://www.cs.waikato.ac.nz/ml/weka/

	TP(R)	FP	P	F	ROC	TP Incr
J48	0.42	0.16	0.40	0.40	0.63	-23%
RandomForest	0.45	0.23	0.46	0.38	0.72	-17%
SMO	**0.54**	**0.14**	**0.53**	**0.52**	**0.80**	**0% (Ref)**
BN	0.48	0.14	0.47	0.47	0.78	-10%
MNB	0.53	0.18	0.54	0.47	0.85	-2%
KNN (k=3)	0.53	0.16	0.53	0.51	0.77	-2%

TAB. 1 – *Résultats de classification sur les données initiales.*

	TP(R)	FP	P	F	ROC	Nbr. var.	TP Incr
CHI+	0.52	0.17	0.51	0.47	0.80	282	-4%
CBF	0.47	0.21	0.44	0.41	0.75	37	-13%
PCA (50% vr.)	0.47	0.18	0.47	0.44	0.77	483	-13%
RLF	0.52	0.16	0.53	0.48	0.81	937	-4%
FMC	**0.99**	**0.003**	**0.99**	**0.99**	**1**	**262/cl**	**+90%**

TAB. 2 – *Résultats de classification après la sélection de variables (classifieur BN).*

n'est clairement pas exploitable dans un contexte opérationnel d'évaluation de brevets, comme celui de QUAERO, en raison de la grande confusion entre les classes, mettant ainsi en évidence son incapacité intrinsèque à faire face à l'effet d'attraction des plus grandes classes. Chaque fois qu'une méthode usuelle de sélection de variables est appliquée en association avec les méthodes de classification les meilleures dans notre contexte, son exploitation altère légèrement la qualité des résultats, comme il est indiqué dans le tableau 2. Le tableau 2 souligne également que la réduction du nombre de variables par la méthode FMC est similaire à CHI+ (en termes de variables actives ; voir la section 2 pour plus de détails) mais que son exploitation stimule les performances des méthodes de classification, et en particulier celles des méthodes bayésiennes (tableau 3), conduisant à des résultats de classification impressionnants dans un contexte de classification très complexe : précision de 0,987%, soit 94 données mal classées parmi un total de 7252 avec la méthode BN. Les résultats présentés dans le tableau 4 illustrent plus précisément l'efficacité de la procédure de contraste F-max qui agit sur les descriptions des données (eq. 6). Dans les expériences relatives à ce tableau, le contraste est appliqué individuellement sur les variables extraites par chaque méthode de sélection et, dans une deuxième étape, un

	TP(R)	FP	P	F	ROC	TP Incr
J48	0.80	0.05	0.79	0.79	0.92	+48%
RandomForest	0.76	0.09	0.79	0.73	0.96	+40%
SMO	0.92	0.03	0.92	0.91	0.98	+70%
BN	**0.99**	**0.003**	**0.99**	**0.99**	**1**	**+90%**
MNB	0.92	0.03	0.92	0.92	0.99	+71%
KNN (k=3)	0.66	0.14	0.71	0.63	0.85	+22%

TAB. 3 – *Résultats de classification après la sélection de variables FMC.*

classifieur BN est appliqué sur les données résultantes contrastées. Les résultats montrent que, quel que soit le type de méthode de sélection de variable utilisé, les performances de classification qui en résultent sont renforcées chaque fois que le contraste F-max est appliqué en aval de la sélection. L'augmentation moyenne de performance est de 44%. Le tableau 5 illustre finalement les capacités de l'approche FMC à faire face efficacement aux problèmes de déséquilibre et de similitude des classes. L'examen des variations des taux TP (surtout dans les petites classes) dans ce dernier tableau montre que l'effet d'attraction de données des plus grandes classes, qui se produit à un niveau élevé dans le cas de l'exploitation des données originales, est pratiquement systématiquement surmonté chaque fois que l'approche FMC est exploitée. La capacité de l'approche à corriger un déséquilibre de classes est également clairement mise en évidence par la répartition homogène des variables actives dans les classes, ceci malgré des tailles très hétérogènes de classe. Le résumé des résultats sur les 3 ensembles de données

	TP(R)	FP	P	F	ROC	Nbr. var.	TP Incr
CHI+	0.79	0.08	0.82	0.78	0.98	282	+46%
CBF	0.63	0.15	0.69	0.59	0.90	37	+16%
PCA (50% vr.)	0.71	0.11	0.73	0.67	0.53	483	+31%
RLF	0.79	0.08	0.81	0.78	0.98	937	+46%
FMC	**0.99**	**0.003**	**0.99**	**0.99**	**1**	**262/cl**	**+90%**

TAB. 4 – *Résultats de classification avec différentes méthodes de sélection de variables et contraste F-max (classifieur BN).*

complémentaires est présenté dans les tableaux 6- 8. Il soulignent que la méthode FMC peut améliorer très significativement les performances des classifieurs dans différents types de cas. Comme dans le contexte de notre expérience précédente (brevets), les meilleure performances sont obtenues par l'exploitation de la méthode FMC en combinaison avec les classifieurs bayésiens MNB et BN. Le tableau 7 présente les résultats comparatifs d'une telle combinaison. Il met en évidence le fait que la méthode FMC est particulièrement efficace pour augmenter les performances des classifieurs dès lors que la complexité de la tâche de classification devient plus élevée en raison d'un nombre croissant de classes (corpus AMZ). Le tableau 8 fournit des informations générales sur les données et sur le comportement de la méthode de sélection FMC. Il illustre la diminution significative de la complexité de classification obtenue avec FMC en raison de la réduction du nombre de variables à gérer, ainsi que la diminution concomitante des données mal classées. Il souligne également le temps de calcul très modéré de la méthode (le calcul est effectué sur Linux avec un ordinateur portable équipé d'un processeur Intel® Pentium® B970 2.3Ghz et avec une mémoire de 8Go.) Sur ces ensembles de données, des remarques similaires à celles mentionnées pour l'ensemble des données-brevets peuvent être faites au sujet de la faible efficacité des méthodes usuelles de sélection de variables et des méthodes de ré-échantillonnage. Le tableau 8 montre également que la valeur du facteur d'amplification du contraste, qui est exploité pour obtenir les meilleures performances, peut varier au fil des expériences (de 1 à 4, dans ce dernier contexte). Cependant, l'on peut observer qu'en prenant une valeur fixe pour ce facteur, par exemple la plus élevée (ici 4), l'on ne dégrade pas les résultats. Ce choix représente donc une bonne alternative pour faire face au problème de paramétrage.

Etiquette classe	Taille	Var. sélect.	% TP FMC	% TP avant
a61k31	2533	223	**1**	0.79
a61k33	60	276	**0.95**	0.02
a61k35	459	262	**0.99**	0.31
a61k36	212	278	**0.95**	0.23
a61k38	1110	237	**1**	0.44
a61k39	1141	240	**0.99**	0.65
a61k41	22	225	**0.24**	0
a61k45	304	275	**0.98**	0.09
a61k47	304	278	**0.99**	0.21
a61k48	140	265	**0.98**	0.12
a61k49	90	302	**0.93**	0.26
a61k51	78	251	**0.98**	0.26
a61k6	47	270	**0.82**	0.04
a61k8	87	292	**0.98**	0.02
a61k9	759	250	**1**	0.45

TAB. 5 – *Caractéristiques/classe avant et après sélection FMC (classifieur BN).*

Trade	**Grain**	**Ship**	**Acq**
6.35 tariff	5.60 agricultur	6.59 ship	5.11 common
5.49 trade	5.44 farmer	6.51 strike	4.97 complet
5.04 practic	5.33 winter	6.41 worker	4.83 file
4.86 impos	5.15 certif	5.79 handl	4.65 subject
4.78 sanction	4.99 land	5.16 flag	4.61 tender
Learn	**Money-fx**	**Interest**	**Crude**
7.57 net	6.13 currenc	5.95 rate	6.99 oil
7.24 loss	5.55 dollar	5.85 prime	5.20 ceil
6.78 profit	5.52 germani	5.12 point	4.94 post
6.19 prior	5.49 shortag	5.10 percentag	4.86 quota
5.97 split	5.16 stabil	4.95 surpris	4.83 crude

TAB. 6 – *Liste des variables (lemmes) de contraste élevé pour les 8 classes du corpus REUTER8.*

		TP (R)	FP	P	F	ROC	TP Incr.
Reuters8 (R8)	-	0.937	0.02	0.942	0.938	0.984	
	FMC	**0.998**	**0.001**	**0.998**	**0.998**	**1**	**+6%**
Reuters52 (R52)	-	0.91	0.01	0.909	0.903	0.985	
	FMC	**0.99**	**0.001**	**0.99**	**0.99**	**0.999**	**+10%**
Amazon	-	0.748	0.05	0.782	0.748	0.981	
	FMC	**0.998**	**0.001**	**0.998**	**0.998**	**1**	**+33%**

TAB. 7 – *Résultats de classifications après sélection de variables FMC (classifieur MNB/BN).*

Les 5 variables (lemmes) les plus contrastées des 8 classes issues du corpus Reuter8 sont présentées dans le tableau 6. Le fait que les grandes lignes des thématiques couvertes par les classes puissent être clairement mises en évidence de cette manière illustre bien les capacités d'extraction de sujets de la méthode FMC. Enfin, l'obtention de très bonnes performances en combinant l'approche de sélection de variables FMC avec une méthode de classification comme MNB est un réel avantage pour l'exploitation à grande échelle, sachant que la méthode MNB a des capacités incrémentales et que les deux méthodes ont des temps de calcul faibles.

	R8	R52	AMZ
Nbr. classes	8	52	50
Nbr. données	7674	9100	1500
Nbr. variables	3497	7369	10000
Nbr. var. sélect.	1186	2617	3318
Var. activ./classe (moy.)	268.5	156.05	761.32
Facteur d'amplification	4	2	1
Mal classés (Std)	373	816	378
Mal classés (FMC)	**19**	**91**	**3**
Temps calcul (s)	1	3	1.6

TAB. 8 – *Informations sur les données et résultats complémentaires après sélection de variables FMC (classifieur MNB/BN).*

4.3 Conclusion

Notre objectif principal était de développer une méthode efficace de sélection et de contraste de variables qui pourrait permettre de surmonter les problèmes habituels liés à la classification supervisée de gros volumes de données textuelles. Ces problèmes sont liés à des classes déséquilibrées avec un degré élevé de similitude entre elles, hébergeant des données fortement multidimensionnelles et bruitées. Pour ce faire, nous avons proposé d'adapter une métrique développée récemment dans le cadre non supervisé au contexte de la classification supervisée. Grâce à diverses expériences sur de grands ensembles de données textuelles, nous avons illustré de nombreux avantages de notre approche, et surtout sa grande efficacité pour améliorer les performances des classifieurs dans un tel contexte, tout en mettant l'accent sur les classifieurs les plus flexibles et les moins gourmands en temps de calcul, comme les classifieurs bayésiens. Un autre avantage de cette méthode est qu'il s'agit d'une approche sans paramètre qui s'appuie sur un schéma simple d'extraction de variables ; elle peut donc être utilisée dans de nombreux contextes, comme dans ceux de l'apprentissage incrémental ou semi-supervisé, ou encore, dans celui de l'apprentissage numérique en général. Une autre perspective intéressante serait d'adapter cette technique au domaine de l'exploration de textes afin d'enrichir des ontologies et des lexiques grâce à l'exploitation à grande échelle des corpus existants.

Remerciements Ce travail a été réalisé dans le cadre du programme QUAERO [5] soutenu par OSEO [6], Agence française de développement de la recherche.

5. http://www.quaero.org
6. http://www.oseo.fr/

Références

Aha, D., D. Kibler, et M. Albert (1991). Instance-based learning algorithms. *Machine learning 6*, 37–66.

Bache, K. et M. Lichman (2013). Uci machine learning repository (http://archive.ics.uci.edu/ml). : University of california, school of information and computer science, irvine, ca, usa.

Bolón-Canedo, V., N. Sánchez-Maroño, et A. Alonso-Betanzos (2012). A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, 1–37.

Breiman, L. (2001). Random forests. *Machine learning 45(1)*, 5–32.

Breiman, L., J. Friedman, R. Olshen, et C. Stone (1984). Classification and regression trees. Technical report, Wadsworth International Group, Belmont, CA, USA.

Chawla, N., K. Bowyer, L. Hall, et W. Kegelmeyer (2002). Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research 16*, 321–357.

Dash, M. et H. Liu (2003). Consistency-based search in feature selection. *Artificial Intelligence 151(1)*, 155–176.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research 3*, 1289–1305.

Good, P. (2006). *Resampling methods*. Ed. Birkhauser.

Guyon, I. et A. Elisseeff (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research 3*, 1157–1182.

Guyon, I., J. Weston, S. Barnhill, et V. Vapnik (2002). Gene selection for cancer classification using support vector machines. *Machine Learning 46(1)*, 389–422.

Hajlaoui, K., P. Cuxac, J.-C. Lamirel, et C. FranÃ§ois (2012). Enhancing patent expertise through automatic matching with scientific papers. In J.-G. Ganascia, P. Lenca, et J.-M. Petit (Eds.), *Discovery Science*, Volume 7569 of *Lecture Notes in Computer Science*, pp. 299–312. Springer Berlin Heidelberg.

Hall, M. et L. Smith (1999). Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference*, pp. 235–239.

Kohavi, R. et G. John (1997). Wrappers for feature subset selection. *Artificial Intelligence 97(1-2)*, 273–324.

Konokenko, I. (1995). Learning to filter netnews. In *Proceedings of European Conference on Machine Learning*, pp. 331–339.

Ladha, L. et T. Deepa (2011). Feature selection methods and algorithms. *International Journal on Computer Science and Engineering 3(5)*, 1787–1797.

Lallich, S. et R. Rakotomalala (2000). Fast feature selection using partial correlation for multi-valued attributes. In D. A. Zighed, J. Komorowski, et J. Å»ytkow (Eds.), *Principles of Data Mining and Knowledge Discovery*, Number 1910 in Lecture Notes in Computer Science, pp. 221–231. Springer Berlin Heidelberg.

Lamirel, J., S. Al Shehabi, C. François, et M. Hoffmann (2004). New classification quality

estimators for analysis of documentary information: application to patent analysis and web mapping. *Scientometrics 60(3)*.

Lamirel, J., M. Ghribi, et P. Cuxac (2010). Unsupervised recall and precision measures: a step towards new efficient clustering quality indexes. In *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010, Paris, France)*.

Lamirel, J. et A. Ta (2008). Combination of hyperbolic visualization and graph-based approach for organizing data analysis results: an application to social network analysis. In *Proceedings ofthe 4th International Conference on Webometrics, Informetrics and Scientometrics and 9th COLLNET Meetings, Berlin, Germany*.

Pearson, K. (1901). On lines an planes of closetst fit to systems of points in space. *Philosophical Magazine 2(11)*, 559–572.

Platt, J. (1999). Advances in kernel methods. Chapter Fast training of support vector machines using sequential minimal optimization, pp. 185–208. Cambridge, MA, USA: MIT Press.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Salton, G. (1971). *Automatic processing of foreign language documents*. Englewood Clifs, NJ, USA: Prentice-Hill.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Witten, I. et E. Frank (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Yu, L. et H. Liu (2003). Feature selection for high-dimensional data: a fast correlation-based filter solution. In *Proceedings of ICML 03*, Washington DC, USA, pp. 856–863.

Summary

Feature maximization is a cluster quality metric which favors clusters with maximum feature representation as regard to their associated data. In this paper we go one step further showing that a straightforward adaptation of such metric can provide a highly efficient feature selection and feature contrasting model in the context of supervised classification. The method is experienced on different types of textual datasets. The paper illustrates that the proposed method provides a clear performance increase in all the studied cases even when a single bag of words model is exploited for data description. We more especially show that this technique can enhance the performance of classification methods whilst very signifcantly outperforming (+90%) the state-of-the art feature selection techniques in the case of the classification of unbalanced, highly multidimensional and noisy textual data, with a high degree of similarity between the classes.