Mesurer la proximité entre corpus par de nouveaux méta-descripteurs

Flavien Bouillot, Pascal Poncelet, Mathieu Roche

Mesurer la proximité entre corpus par de nouveaux méta-descripteurs

Flavien Bouillot[*,**] — **Pascal Poncelet**[*] — **Mathieu Roche**[***]

[*] *LIRMM, Univ. Montpellier 2, CNRS – France*
[**] *ITESOFT, Aimargues – France*
[***] *TETIS, Cirad, Irstea, AgroParisTech – France*

prénom.nom@lirmm.fr

RÉSUMÉ. *Devant le nombre d'algorithmes de classification existants, trouver l'algorithme qui sera le plus adapté pour classer un corpus de documents est une tâche difficile. La méta-classification apparaît aujourd'hui très utile pour aider à déterminer, en fonction des expériences passées, quel devrait être l'algorithme le plus pertinent par rapport à notre corpus. L'idée sous jacente est que "si un algorithme s'est montré particulièrement adapté pour un corpus, il devrait avoir le même comportement sur un corpus assez similaire". Dans cet article, nous proposons de nouveaux méta-descripteurs reposant sur les notions de similarités pour améliorer l'étape de méta-classification. Les expérimentations menées sur différents jeux de données réelles montrent la pertinence de nos nouveaux descripteurs.*

ABSTRACT. *Given the number of existing classification algorithms, finding the most appropriate for classifying a new corpus is a difficult task. Meta-classification appears today very useful to help to determine, by using past experiences, what should be the most suitable algorithm compared to our corpus. The underlying idea is that "if an algorithm was particularly suitable for a corpus, it should have the same behavior on a quite similar corpus.". In this paper, we propose new meta-descriptors based on the concept of similarity to improve the meta-classification step. Conducted experiments on real dataset show the relevance of our new meta-descriptors.*

MOTS-CLÉS : *méta-classification, méta-descripteurs, similarité.*

KEYWORDS: *meta-classification, meta-features, similarity.*

1. Introduction

Devant la grande diversité d'algorithmes de classification disponibles, déterminer quel sera l'algorithme le plus approprié face à un nouveau corpus est difficile (Kalousis *et al.*, 2004). Cette difficulté est mise en évidence dans le théorème du *No Free Lunch* (Wolpert et Macready, 1997) qui souligne que si un classifieur A est meilleur qu'un classifieur B pour un corpus donné, alors il existe autant de cas où B sera meilleur. En d'autres termes, ce théorème montre qu'il n'est pas possible de déterminer un classifieur qui fonctionne mieux que les autres indépendamment du problème donné. Outre le choix du classifieur, un problème sous jacent apparait : comment détecter les paramètres les plus adaptés pour un algorithme (Pavón *et al.*, 2009) ?
Traditionnellement pour déterminer de façon automatique les candidats les plus pertinents, il existe deux grandes catégories de travaux. La première regroupe les approches qui essaient de converger vers le meilleur classifieur par itérations successives (approche "*grid search*", programmation génétique). Ce principe est, par exemple, souvent utilisé pour la détection de paramètres (Cristianini *et al.*, 1998). La seconde regroupe les approches qui utilisent les expériences passées pour prédire le futur (Ali et Smith-Miles, 2006). On parle alors de meta-learning ou de méta-classification. Le terme méta-learning a été utilisé la première fois dans (Aha, 1992) mais les approches de méta-classification trouvent leurs inspirations dans le problème de sélection d'un algorithme initialement formalisé par (Rice, 1976). La méta-classification vise à déterminer l'algorithme le plus approprié pour traiter un nouveau problème de classification sur la base des expériences passées (Smith-Miles, 2009). Pour cela, les problèmes étudiés sont décrits au moyen de méta-descripteurs. Plus précisément il s'agit de trouver les descripteurs significatifs des différents corpus utilisés. Par la suite, par application d'un méta-classifieur, il s'agit de prédire les performances des différents algorithmes pour un nouveau problème en recherchant dans les expériences passées celles pour lesquelles les corpus étaient les plus similaires et qui ont eu de bons résultats de classification. Si un classifieur a été performant pour ces corpus, il a de grandes chances de l'être pour un nouveau corpus similaire. L'objectif de cet article est de proposer de nouveaux méta-descripteurs pour décrire les corpus en se reposant sur les notions de similarité. Ces derniers seront ensuite intégrés dans une étape de méta-classification.

Cet article est organisé de la manière suivante. Pour commencer nous présentons, Section 2, un aperçu des différentes propositions de la littérature pour décrire un corpus. Section 3, nous présentons de nouveaux descripteurs adaptés à la représentation de corpus textuel et nous étudions les performances de notre proposition au travers de différentes expérimentations menées sur des jeux de données réelles dans la Section 4. Enfin, Section 5, nous concluons en présentant des travaux futurs.

2. État de l'art sur les méta-descripteurs

Comme nous l'avons précisé en introduction, une grande partie des travaux portant sur la méta-classification s'intéresse à la manière de décrire les jeux de données de la manière la plus précise possible. Ainsi, traditionnellement les premiers méta-descripteurs proposés sont des variables quantitatives calculées sur les jeux de données (e.g. le nombre d'exemples d'apprentissage, le nombre de classes...) (Brazdil *et al.*, 2003 ; Kalousis *et al.*, 2004).

Il est possible de regrouper ces méta-descripteurs quantitatifs en trois familles (Bhatt *et al.*, 2012) :

– Les descripteurs simples qui regroupent les propriétés accessibles rapidement : nombre de classes, nombre d'observations, nombre d'attributs ;

– Les descripteurs statistiques qui reposent sur des méthodes d'analyse statistique : moyenne, variance, matrice de covariance, coefficient d'aplatissement de Pearson, coefficient de dissymétrie ;

– Les descripteurs issus de la théorie de l'information : Khi-2, information Gain et entropie des attributs et des classes.

Généralement les approches empruntent des méta-descripteurs issus des différentes familles. Par exemple dans (Bhatt *et al.*, 2013), les méta-descripteurs utilisés sont à la fois des mesures simples et des mesures statistiques. Ces variables quantitatives peuvent parfois être transformées en variables qualitatives comme dans (Molina *et al.*, 2012). Enfin, dans un contexte plus spécifique de classification de documents, des spécificités linguistiques peuvent être associées. Ainsi, (Lam et Lai, 2001) et (Furdík *et al.*, 2008) utilisent des méta-descripteurs tels que le nombre d'exemples, le nombre de descripteurs linguistiques moyen par document par catégorie, le nombre de descripteurs linguistiques moyen ayant un poids supérieur à un seuil donné ou encore le Gain d'Information moyen pour les meilleurs t descripteurs linguistiques de chaque classe.

D'autres approches ont été proposées et utilisent les résultats d'algorithmes pour décrire les corpus. Dans (Peng *et al.*, 2002), les auteurs utilisent comme méta-descripteurs les propriétés d'un arbre de décision généré sur les jeux de données (largeur de l'arbre, profondeur de l'arbre). (Sun et Pfahringer, 2013) proposent des méta-descripteurs basés sur la relation entre algorithmes. Ils suggèrent un nouveau modèle de génération de méta-descripteurs en comparant les algorithmes 2 à 2 et en obtenant des règles de préférences. Une autre approche qui utilise aussi les performances comme méta-descripteurs a été utilisée dans (Leite et Brazdil, 2005). Les auteurs utilisent une approche fondée sur les courbes d'apprentissage (learning curves). Enfin, nous pouvons citer un autre type de méta-descripteurs qui s'appuient sur le temps de calcul constaté pour la mesures des divers méta-descripteurs (Reif *et al.*, 2011).

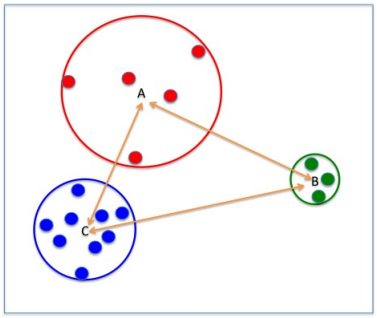

Figure 1. *Exemple de corpus projetés dans un espace euclidien à 2 dimensions*

Dans (Pfahringer *et al.*, 2000), les auteurs ont proposé la notion de *landmarkers*. Le principe général est d'utiliser les performances d'algorithmes de classification pour caractériser un jeu de données et repose sur l'hypothèse que si deux jeux de données qui ont des caractéristiques similaires ont des résultats similaires pour un ensemble d'algorithmes de classification donné alors deux jeux de données qui ont des résultats similaires pour en ensemble d'algorithme donné auront des caractéristiques identiques. Concrètement les méta-descripteurs sont représentés par les scores obtenus en appliquant des algorithmes rapides de classification issus de différentes familles. Pour ce dernier point l'idée est que les algorithmes construits sur les mêmes fondements auront des comportements et des biais similaires. L'intérêt d'avoir des algorithmes rapides est bien entendu d'éviter de lancer une classification complexe sur le corpus mais de voir le comportement général de l'algorithme.

3. Une nouvelle façon de décrire un corpus

Dans la suite de cet article, nous proposons de décrire un corpus en fonction de sa représentation dans l'espace euclidien. L'hypothèse sous jacente est qu'en positionnant les composantes d'un corpus, les uns par rapport aux autres, il est possible d'obtenir un schéma, une empreinte, qui soit caractéristique du corpus.

Dans un processus de classification, deux composants (classes ou documents) sont similaires lorsque le vocabulaire utilisé dans les deux composants est similaire. L'objectif de nos descripteurs est de tenir compte de ces similarités pour mieux caractériser le corpus. Nous considérons donc *la similarité inter-classes* et *la similarité intra-classe* afin de mieux caractériser comment les différents éléments se projettent dans l'espace euclidien. La Figure 1 illustre trois classes A, B et C. Comme nous pouvons le constater les classes A et C sont plus proches par rapport à B (similarité inter-classes) et la similarité intra-classe fait apparaître que les documents de la classes B sont plus similaires (puisque moins éloignés) que ceux de la classe A.

De nombreuses mesures de similarité existent dans la littérature. Une étude comparative est proposée par exemple dans (Wajeed et Adilakshmi, 2011) et (Lin *et al.*, 2013). Certaines mesures sont applicables uniquement à des attributs binaires (Jaccard, Dice) alors que d'autres sont applicables sur des attributs non-binaires (Distance Euclidienne, Cosinus, Tanimoto). Nous nous focalisons dans la suite sur le coefficient de Tanimoto qui présente l'avantage d'être borné entre 0 et 1 et qui permet de prendre en considération la fréquence des termes et non uniquement leur absence ou présence. Nous précisons néanmoins que le principe proposée est applicable avec toute autre mesure de similarité sur des attributs non-binaires (notamment le cosinus).

Pour une classe (ou un document), le coefficient de Tanimoto compare le poids de la somme des termes communs à la somme des poids des termes qui sont présents dans l'une des deux classes (ou documents), mais qui ne sont pas des termes communs aux deux. La définition formelle est :

$Coefficient de Tanimoto : S_J(d_1, d_2) = \frac{\sum_{k=1}^{N}(w_{k1} \times w_{k2})}{\sum_{k=1}^{N} w_{k1}^2 + \sum_{k=1}^{N} w_{k2}^2 - \sum_{k=1}^{N}(w_{k1} \times w_{k2})}$
où w_{k1} (respectivement w_{k2}) correspond au poids (i.e. la fréquence) du k^{me} terme pour la classe 1 (respectivement la classe 2). A noter que nous utilisons les termes dans leurs formes fléchies comme descripteurs mais que notre méthode est applicable avec les autres types de descripteurs (lemmes, racines, grammes de mots ou de caractères).

De manière à illustrer l'application du coefficient de Tanimoto pour déterminer les similarités intra-classe et inter-classes, considérons un corpus exemple (Table 1) qui regrouperait des documents liés au monde animal en fonction de leur famille (issue de la classification des espèces). Ce corpus exemple est composé :

– de classes plus similaires que d'autres (les familles Felidae et Canidae appartiennent toutes deux à l'ordre Carnivora) ;

– de classes aux vocabulaires plus larges que d'autres.

La représentation sac de mots des classes Canidae et Felidae est :

– $V_{Canidae}$:{animal (5), balle (1), cage (1), carnivore (2), chacal (1), chien (3), chienne (1), chiot (3), compagnie (4), croquette (1), griffe (1), laisse (1), loup (1), mammifère (2), medor (1), milou (1), pif (1), rantanplan (1), renard (1), spa (1), toutou (3)}

– $V_{Felidae}$:{animal (5), balle (1), cage (1), carnivore (2), chat (2), compagnie (4), croquette (1), griffe (1), laisse (1), lynx (1), mammifère (2), matou (4), minet (1), minou (7), puma (1), shereKhan (1), spa (1), tigre (1)}

et la similarité entre les classes Canidae et Felidae est : $S_J(V_{Canidae}, V_{Felidae}) = \frac{\sum_{k=1}^{N}(w_{k1} \times w_{k2})}{\sum_{k=1}^{N} w_{k1}^2 + \sum_{k=1}^{N} w_{k2}^2 - \sum_{k=1}^{N}(w_{k1} \times w_{k2})} = \frac{55}{90+129-55} = 0.34$

Classe	Documents	descripteurs linguistiques
Bovidae	bov_1	animal-compagnie-boeuf-génisse-vachette-yack-yack-yack
	bov_2	animal-boeuf-taureau
	bov_3	animal-yack
	bov_4	animal-herbe-marguerite-ruminer-train-vachette
	bov_5	animal-corne-mamifère-peau-yack
Canidae	can_1	animal-compagnie-chien-chienne-chiot-toutou
	can_2	animal-compagnie-chiot-chiot-toutou-toutou
	can_3	animal-compagnie-milou-medor-rantanplan-pif
	can_4	animal-carnivore-chacal-mammifère-renard
	can_5	balle-chien-croquette-laisse
	can_6	cage-carnivore-griffe-loup-mammifère
	can_7	animal-chien-compagnie-spa
Felidae	fel_1	animal-chat-compagnie-matou-matou-minet
	fel_2	animal-compagnie-matou-minou-minou-minou
	fel_3	animal-compagnie-matou-minou-minou-minou
	fel_4	animal-carnivore-lynx-mammifère-puma
	fel_5	balle-croquette-laisse-minou
	fel_6	cage-carnivore-griffe-shereKhan-tigre
	fel_7	animal-chat-compagnie-mammifère-spa
Ornithorhynchidae	ort_1	animal-australie-carnivore-mammifère-ornithorynque-toto
	ort_2	australie-hexley-mammifère-ornithorynque
	ort_3	animal-australie-carnivore-mammifère-ornithorynque-poil-venin
	ort_4	animal-australie-bec-carnivore-eau-mammifère-œuf-ornithorynque-palme
Anatidae	ana_1	cygne-cygne-eau-mare-vilain
	ana_2	animal-canard-confit-foie-magret
	ana_3	aile-animal-bec-œuf-oie-palme-plume
	ana_4	canardo-daffy-donald-saturnin
	ana_5	animal-canardeau-cancane-cane-caneton-canette

Tableau 1. *Corpus exemple*

	Anatidae	Bovidae	Canidae	Felidae	Ornithorhynchidae
Anatidae		0.17	0.14	0.10	0.14
Bovidae	0.17		0.22	0.17	0.12
Canidae	0.14	0.22		0.34	0.21
Felidae	0.10	0.17	0.34		0.17
Ornithorhynchidae	0.14	0.12	0.21	0.17	

Tableau 2. *Calcul des similarités inter-classes*

Bien entendu, le nombre de similarités à calculer dépend du nombre de classes du corpus. Ainsi, pour un corpus composé de |C| classes, le nombre de similarités inter-classes à calculer est égal à $\frac{|C|(|C|-1)}{2}$.

De la même manière nous pouvons calculer les similarités intra-classes. Par exemple pour la classe Felidae, nous calculons les similarités S_J pour chaque paire de documents de la classe ($S_J(fel\_1, fel\_2)$, $S_J(fel\_1, fel\_3)$...) dans le tableau 3.

Nous pouvons remarquer que $S_J(V_{fel\_2}, V_{fel\_3}) = 1$ (les documents sont identiques) et $S_J(V_{fel\_2}, V_{fel\_6}) = 0$ (les documents ne partagent aucun vocabulaire commun).

$S_J(x,y)$	fel_1	fel_2	fel_3	fel_4	fel_5	fel_6	fel_7
fel_1		0.25	0.25	0.08	0	0	0.3
fel_2	0.25		1	0.063	0.23	0	0.13
fel_3	0.25	1		0.063	0.23	0	0.13
fel_4	0.083	0.063	0.063		0	0.11	0.25
fel_5	0	0.23	0.23	0		0	0
fel_6	0	0	0	0.11	0		0
fel_7	0.3	0.13	0.13	0.25	0	0	

Tableau 3. *Calcul des similarités intra-classe pour la classe Felidae*

Pour une classe j donnée, le nombre de similarités intra-classes à calculer va dépendre du nombre de documents qu'il possède. Ainsi pour d documents, nous avons $\frac{|d|(|d|-1)}{2}$ calculs à effectuer.

Pour résumer, tout corpus de |*C*| classes peut donc être décrit par des ensembles de similarités indépendants :

– *un* ensemble de similarités inter-classes ;

– |*C*| ensembles de similarités intra-classes.

Malheureusement l'utilisation de ces similarités comme méta-descripteurs n'est pas possible. Cette représentation implique qu'en cas d'ajout ou de suppression d'un document ou d'une classe, le nombre de similarités calculé sera différent, rendant la comparaison de deux corpus impossible.

Pour résumer un nombre variable de valeurs en un nombre fini de valeurs, nous proposons d'utiliser un ensemble de statistiques descriptives simples telles que la moyenne, la médiane ou le maximum.

Considérons les 10 similarités inter-classes calculées dans notre exemple, nous pouvons les résumer en utilisant 4 statistiques descriptives (moyenne, variance, minimum et maximum). Nous obtenons ainsi 4 valeurs (0.18[1], 0.005[2], 0.10[3], 0.34[4]).

Ces valeurs représentent la similarité moyenne entre les classes du corpus, la variance des similarités entre les classes ou encore la similarité minimum et maximum entre les classes du corpus et permettent de savoir, par exemple, si les classes sont en moyennes proches les unes des autres (moyenne des similarités inter-classes), si

1. moyenne(0.17, 0.14, 0.10, 0.14, 0.22, 0.17, 0.12, 0.34, 0.21, 0.17)=0.18
2. variance(0.17, 0.14, 0.10, 0.14, 0.22, 0.17, 0.12, 0.34, 0.21, 0.17)=0.005
3. minimum(0.17, 0.14, 0.10, 0.14, 0.22, 0.17, 0.12, 0.34, 0.21, 0.17)=0.10
4. maximum(0.17, 0.14, 0.10, 0.14, 0.22, 0.17, 0.12, 0.34, 0.21, 0.17)=0.34

Classe	Moyenne	Variance	Minimum	Maximum
Bovidae	0.20	0.001	0.10	0.40
Canidae	0.13	0.024	0	0.60
Felidae	0.15	0.05	0	1
Ornithorhynchidae	0.45	0.0125	0.30	0.63
Anatidae	0.03	0.002	0	0.10

Tableau 4. *Résumé des similarités intra-classe*

les similarités entre classes sont identiques (variance des similarités inter-classes) ou encore quel est l'écart minimum ou maximum entre classes (minimum et maximum des similarités inter-classes).

En utilisant d statistiques descriptives, nous pouvons résumer les similarités inter-classes à l'aide de d valeurs numériques indépendamment du nombre de classes du corpus. Il devient alors possible de comparer 2 ensembles de similarités inter-classes en comparant les d valeurs statistiques.

Concernant l'agrégation des similarités intra-classes, il est nécessaire de se rappeler que les similarités intra-classes sont calculées pour une classe donnée indépendamment des autres classes (un ensemble de valeurs pour la classe Felidae puis un autre ensemble pour la classe Bovidae...).

Considérer les similarités intra-classes comme un seul et même ensemble reviendrait à faire abstraction du comportement spécifique de chaque classe. Or certaines classes sont composées de documents composés de vocabulaire différent, d'autres de documents proches et le plus souvent elles sont un mélange des deux. C'est cette hétérogénéité que nous souhaitons mesurer. L'objectif est d'agréger ***n*** ensembles composés d'un nombre variable de valeurs (***n*** correspondant au nombre de classes, le nombre de valeurs dépendant du nombre de documents de la classe) en un nombre fini d'ensembles composés d'un nombre fini de valeurs. Pour cela, nous proposons une approche en deux étapes :

1) Agréger indépendamment les ***n*** ensembles composés d'un nombre variable de valeurs en ***n*** ensembles composés d'un nombre fini de valeurs via des statistiques descriptives.

2) Agréger les ***n*** ensembles obtenus à l'étape 1 en un nombre fini d'ensembles en utilisant les mêmes statistiques descriptives.

Pour notre exemple nous mesurons tout d'abord la similarité moyenne au sein de chaque classe, la variance au sein de chaque classe ou encore le minimum et le maximum des similarités entre documents au sein de chaque classe (étape 1). Nous obtenons ainsi 5 ensembles de 4 valeurs statistiques (5 classes et 4 descripteurs statistiques) que nous résumons dans le tableau 4.

	Moyenne	Variance	Minimum	Maximum
Inter-Classe	0.18	0.005	0.10	0.34
$Intra - Classe^{moyenne}$	0.19	0.02	0.03	0.45
$Intra - Classe^{variance}$	0.02	0	0	0.05
$Intra - Classe^{minimum}$	0.08	0.02	0	0.30
$Intra - Classe^{maximum}$	0.55	0.11	0.10	1

Tableau 5. *Résumé des similarités inter-classes et intra-classes*

Nous agrégeons ensuite l'ensemble des moyennes (puis l'ensemble des variances, des minimum et des maximum) en utilisant les mêmes statistiques (étape 2). Par exemple, nous agrégeons les moyennes (0.20, 0.13, 0.15, 0.45, 0.03) en calculant la moyenne des moyennes (0.19), puis la variance des moyennes (0.02), le minimum des moyennes (0.03) et le maximum des moyennes (0.45).

Ainsi indépendamment du nombre de classes et du nombre de documents par classe, il est possible de définir la représentation intra-classe du corpus au moyen de **|D|** ensembles de **|D|** de valeurs, où **|D|** correspond au nombre de descripteurs statistiques choisis pour décrire le corpus. En ajoutant les **|D|** valeurs calculées pour la similarité inter-classes, nous pouvons au final décrire notre corpus au travers de **|D|x|D|+|D|** descripteurs tels que résumé dans le tableau 5.

Prises indépendamment les |D|x|D|+|D| valeurs ne permettent pas d'avoir une vision précise de la représentation d'un corpus mais prises ensembles, elles fournissent une description complète et cohérente du comportement des classes et de leurs compositions. Nous proposons d'utiliser ces |D|x|D|+|D| valeurs comme autant de méta-descripteurs pour décrire un corpus.

Dans nos exemples précédents nous avons retenu 4 statistiques descriptives différentes. Il convient bien entendu de bien définir l'ensemble D des descripteurs utilisés. En effet en en choisissant un nombre trop faible, la description ne sera pas assez précise (Il est possible d'avoir deux moyennes similaires pour des réalités bien différentes), inversement un nombre trop grand de descripteurs peut entraîner un sur-apprentissage. Nous revenons sur ce point dans la section suivante en évaluant différentes valeurs de descripteurs.

4. Expérimentations

Dans cette section, nous évaluons notre proposition par rapport aux autres approches de méta-descripteurs présentées précédemment. Notre objectif est de pouvoir évaluer si nos descripteurs caractérisent bien le corpus et peuvent être utilisés dans une étape de méta-classification.

Corpus	Nb de classes	Nb de documents	Nb de descripteurs	Nb de descripteurs distincts
Dépêches	39	14 701	1 237 264	59 281
Tweets	5	1 186	1 579 374	16 593
Courriers	6	1 273	124 538	23 824

Tableau 6. *Les différents corpus à partir desquels sont générés les 111 corpus uniques*

4.1. *Protocole expérimental*

Les expérimentations ont été menées sur trois corpus différents : un corpus d'échanges de courrier, un corpus de dépêches en anglais (Reuters), un corpus de tweets. Leurs caractéristiques sont résumées dans le Tableau 6. Les descripteurs utilisés sont les termes dans leurs formes fléchies. Nous ne souhaitions pas appliquer de prétraitement différents selon les corpus.

A partir de ces trois corpus très différents nous avons généré 111 corpus uniques (par suppression aléatoire de classes, de documents ou de descripteurs linguistiques).

Sur chacun de ces corpus, nous avons exécuté 8 algorithmes (*naivebayes*, *naivebayesmultinomial*, *complementnaivebayes*, *dmnb*, *libsvm*, *SMO*, *j48* et *LadTree*) à partir du logiciel Weka et nous avons relevé 6 mesures différentes (la Micro Précision, le Micro Rappel, la Micro F-mesure ainsi que la Macro précision, le Macro Rappel et la Macro F-mesure).

Pour comparer notre proposition avec celles de la littérature, nous avons cherché à prédire, pour chacun des 48^5 couples <algorithme de classification, mesure>, le score obtenu en construisant un modèle à partir des 111 corpus. Pour cela nous avons extrait les 111 ensembles de méta-descripteurs à partir des 111 corpus et par application d'un algorithme de régression, nous avons mesuré l'écart entre les scores prédits et les scores réels. Cet écart a été évalué avec le coefficient de corrélation (qui tend vers 1 lorsque les scores prédits sont proches des scores réels, 0 sinon). A noter qu'il n'est pas ici question de la qualité du score en lui-même, mais bien de l'écart entre le score prédit et le score réel (un score faible prédit quand un score faible est réellement observé donne un fort coefficient de corrélation). Afin d'obtenir un résultat robuste et dans la mesure où le nombre de corpus à notre disposition pour, à la fois, construire et évaluer les différents modèles de régression était limité, nous avons utilisé une validation croisée *leave-one-out*.

Pour évaluer la qualité des modèles, nous avons sélectionné un ensemble d'algorithme de régression. Nous avons sélectionné 12 algorithmes appartenant à différentes familles (SVM, Arbre de décision, K plus proche voisin, Régréssion Linéaire...) afin d'éviter d'introduire des biais dans nos analyses en observant des résultats qui seraient liés aux algorithmes de régression et non à nos méta-descripteurs. Pour les

5. 8 algorithmes de classification $\times$ 6 mesures

mêmes raisons, pour isoler les impacts de notre contribution par rapport aux éléments existants, nous avons utilisé les implémentations fournies dans Weka.
Parmi les 12 algorithmes testés, nous avons ensuite sélectionné les 5 algorithmes[6] qui permettaient d'obtenir les meilleurs résultats sur nos données (*IB2*, *KStar*, *LinearRegression*, *M5P*, *SMOreg*).

Il convient maintenant de discuter des méta-descripteurs utilisés, à la fois ceux que nous proposons, mais aussi ceux de la littérature avec lesquels nous nous comparons.

4.2. *Quatre ensembles de méta-descripteurs comparés*

Pour nos expérimentations, nous avons utilisé deux ensembles de descripteurs statistiques. Le premier (que nous nommons *méta-D-14*) est composé de 14 descripteurs statistiques, D_1={le minimum, le maximum, la moyenne, la dispersion, la variance, le 1er décile, le 2ème décile, le 3ème décile, le 4ème décile, le 5ème décile, le 6ème décile, le 7ème décile, le 8ème décile, le 9ème décile}. Ces 14 descripteurs génèrent 210 méta-descripteurs.

Pour notre second ensemble de descripteurs statistiques, nous avons décidé de supprimer les déciles (en conservant néanmoins le 5ème décile qui correspond à la médiane). Nous conservons ainsi 6 descripteurs statistiques, D_2={le minimum, le maximum, la moyenne, la dispersion, la variance, le 5ème décile}, et définissons ainsi 42 méta-descripteurs. Par analogie, nous ferons référence aux 42 méta-descripteurs basés sur ces 6 descripteurs statistiques sous le nom *méta-D-6* dans la suite de cet article.

Parmi les différentes propositions de la littérature, nous avons retenu les approches basées sur les landmarkers (Pfahringer *et al.*, 2000) qui sont considérées comme les plus efficaces pour décrire un jeu de données dans un contexte de méta-classification (Leite et Brazdil, 2010). Nous avons décidé d'utiliser 5 landmarkers qui correspondent aux critères attendus (ils sont rapides et appartiennent des différentes familles) : *complementnaivebayes*, *ib1*, *oner*, *reptree* et *zeror*. Dans la suite de cet article, nous ferons référence aux méta-descripteurs basés sur les landmarkers sous le nom *méta-land*.

Nous avons également utilisé un ensemble de méta-descripteurs composés de statistiques élémentaires, appelé *méta-stat*, calculées sur les corpus : le nombre de classes, le nombre de documents, le nombre de descripteurs linguistiques, le nombre

6. Dans 93% des cas testés, l'un de ces 5 algorithmes retenus donne de meilleurs résultats que les 11 autres.

de descripteurs linguistiques uniques.

Nous présentons dans la section suivante une synthèse des résultats obtenus pour les 4 ensembles de méta-descripteurs.

4.3. *Résultats*

Dans un premier temps nous avons choisi de comparer les résultats obtenus avec les 4 ensembles méta-descripteurs *méta-land*, *méta-stat*, *méta-D-14* et *méta-D-6*. Nous souhaitions savoir si nos propositions nous permettaient d'obtenir de meilleurs résultats que les approches de la littérature retenues. Pour ce faire, nous avons évalué chacun des 48 couples <algorithmes de classification-mesures> à partir des 5 algorithmes de régressions en utilisant les méta-descripteurs *méta-land*. Puis nous avons évalué chacun des 48 couples <algorithmes de classification-mesures> à partir des 5 algorithmes en utilisant les méta-descripteurs *méta-stat*, puis avec les méta-descripteurs *méta-D-14* et *méta-D-6*. Ainsi pour chacun des triplets <algorithme de régression - algorithmes de classification - mesures >, nous avons relevé 4 coefficients de corrélation. Un sous-ensemble de résultats est donné Tableau 7 où les scores maximums ont été signifiés en gras-rouge. L'ensemble le plus performant étant celui nous permettant d'obtenir le coefficient de corrélation le plus important.

Algo	A	Mesure	*méta-D-14*	*méta-D-6*	*méta-land*	*méta-Stat*
KStar	dmnb	macro fscore	**0.97**	0.95	0.96	0.80
KStar	dmnb	macro prec	**0.96**	0.92	0.94	0.77
KStar	dmnb	macro rapp	0.93	0.94	**0.95**	0.81
...	...	...	...	...	...	...
M5P	naivebayes	macro prec	0.94	**0.97**	0.96	0.90
M5P	naivebayes	micro prec	**0.96**	0.92	0.92	0.95
M5P	naivebayes	micro rapp	0.91	0.88	0.94	**0.95**

Tableau 7. *Sous-ensemble des coefficients de corrélation obtenus*

Sur ce sous-ensemble, nous pouvons déjà observer que selon l'algorithme de régression, l'algorithme de classification évalué et la mesure, il est difficile de faire émerger une approche de méta-descripteurs clairement plus performante que les autres (à part *méta-stat* qui possède les plus mauvais scores).

Pour synthétiser les résultats, nous avons mesuré les écarts moyens observés en comparant les ensembles de méta-descripteurs deux à deux. Notre objectif était de voir si, sur l'ensemble des observations, l'un des 4 ensembles de méta-descripteurs nous permet d'obtenir globalement de meilleurs performances dans la mesure où nous venons de démontrer la pertinence de chacun d'eux pour au moins un problème donné. Ainsi nous avons calculé les écarts entre les ensembles *méta-D-14* et *méta-D-6* sur l'ensemble des observations puis nous avons considéré la moyenne des écarts.

Puis nous avons effectué cette comparaison pour l'ensemble de 6 couples et nous présentons les résultats dans le Tableau 8.

Méta-descripteurs 1	Méta-descripteurs 2	Ecart Moyen
méta-D-14	*méta-D-6*	-0.04
méta-D-14	*méta-land*	-0.04
méta-D-14	*méta-stat*	0.09
méta-D-6	*méta-land*	0.00
méta-D-6	*méta-stat*	0.13
méta-land	*méta-stat*	0.14

Tableau 8. *Comparaison des 4 ensembles de méta-descripteurs*

Les résultats nous permettent d'établir une hiérarchie entre les différents ensembles. *méta-D-6* et *méta-land* sont très proches sur la globalité des cas et ils donnent de meilleurs résultats que les méta-descripteurs *méta-D-14* et *méta-stat*. Les méta-descripteurs *méta-D-14* permettent néanmoins d'obtenir de meilleurs résultats que les méta-descripteurs *méta-stat*.

Ces expérimentations nous permettent de conclure que l'utilisation de notre proposition est une alternative pertinente aux méta-descripteurs existants et nous discutons dans la section suivante des avantages et inconvénients des différents ensembles de méta-descripteurs.

5. Conclusions

Dans cet article, nous proposons et expérimentons une façon innovante de décrire un corpus. Nous proposons une nouvelle approche permettant de résumer un corpus en un nombre fini de méta-descripteurs indépendamment du nombre de classes ou de documents par classes, indépendamment du type de documents du corpus, de la langue utilisée. Les nouveaux méta-descripteurs proposés nous permettent d'obtenir des résultats tout à fait comparables aux approches de la littérature. L'approche proposée permet de définir un grand nombre de méta-descripteurs à partir d'un nombre restreint de descripteurs statistiques et il convient de choisir le juste équilibre pour éviter les effets de sur-apprentissage ou de sous-apprentissage. Dans nos expérimentations, nous avons comparé deux ensembles de descripteurs statistiques (composés respectivement de 6 et 14 descripteurs statistiques) en obtenant de meilleures performances avec l'ensemble le plus restreint. La mise en oeuvre de notre approche est aisée puisque qu'elle ne repose que sur le calcul de similarités entre deux documents (ou classes) puis sur l'utilisation de fonctions statistiques simples. Nos méta-descripteurs permettent de comprendre et de visualiser le type de corpus traité (Ai-je des classes très proches sémantiquement parlant ou au contrairement

très disparates ? Les documents au sein de mes classes sont-ils homogènes ou hétérogènes ?). Cette connaissance peut être affinée lors du processus d'extraction des méta-descripteurs. En effet lors des étapes intermédiaires, nous obtenons une vision extrêmement détaillée du comportement de chaque classe avant de les agréger pour obtenir un nombre fini de méta-descripteurs. Cette analyse peut permettre de détecter des comportements anormaux (pourquoi ai-je un document qui se retrouve à l'écart de la majorité des documents de ma classe ? Pourquoi deux classes sont elles si proches sémantiquement alors qu'a priori rien ne le prédisposait ?) ou de modifier le corpus en conséquence (ne serait-il pas préférable de regrouper deux classes sémantiquement trop proches ? ou au contraire scinder une classe en deux ?). Notre approche n'implique pas de partitionner les corpus comme c'est le cas avec les approches basées sur les landmarkers, ce qui nous garantit l'absence de biais introduit par l'utilisation de sous-corpus. De plus, notre approche permet de séparer les problématiques de classification des problématiques de description contrairement aux landmarkers qui se basent sur des problématiques de classification pour traiter une problématique de description.

Nos travaux futurs s'intéressent à la sélection des statistiques descriptives les plus significatives pour décrire le corpus et à étudier l'impact de ces choix (cf les différences de résultats entre *méta-D-14* vs *méta-D-6*).

6. Bibliographie

Aha D. W., « Generalizing from Case Studies A Case Study », *In Proceedings of the Ninth International Conference on Machine Learning*, Morgan Kaufmann, p. 1-10, 1992.

Ali S., Smith-Miles K. A., « A meta-learning approach to automatic kernel selection for support vector machines », *Neurocomputing*, vol. 70, n° 103, p. 173 - 186, 2006. Neural Networks Selected Papers from the 7th Brazilian Symposium on Neural Networks (SBRN '04) 7th Brazilian Symposium on Neural Networks.

Bhatt N., Thakkar A., Ganatra A., « A survey and current research challenges in meta learning approaches based on dataset characteristics », *International Journal of Soft Computing and Engineering*, vol. 2, n° 10, p. 234-247, 2012.

Bhatt N., Thakkar A., Ganatra A., Bhatt N., « Ranking of Classifiers based on Dataset Characteristics using Active Meta Learning », *International Journal of Computer Applications*, vol. 69, n° 20, p. 31-36, May, 2013. Published by Foundation of Computer Science, New York, USA.

Brazdil P. B., Soares C., Da Costa J. P., « Ranking Learning Algorithms Using IBL and Meta-Learning on Accuracy and Time Results », *Mach. Learn.*, vol. 50, n° 3, p. 251-277, March, 2003.

Cristianini N., Campbell C., Shawe-taylor J., « Dynamically Adapting Kernels in Support Vector Machines », *Advances in Neural Information Processing Systems 11*, MIT Press, p. 204-210, 1998.

Furdík K., Paralič J., Tutoky G., « Meta-learning method for automatic selection of algorithms for text classification », *Proc. of the Central European Conference on Information and Intelligent Systems (CECIIS 2008)*, p. 24-26, 2008.

Kalousis A., Gama J. a., Hilario M., « On Data and Algorithms Understanding Inductive Performance », *Mach. Learn.*, vol. 54, n° 3, p. 275-312, March, 2004.

Lam W., Lai K.-Y., « A Meta-learning Approach for Text Categorization », *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, ACM, New York, NY, USA, p. 303-309, 2001.

Leite R., Brazdil P., « Predicting Relative Performance of Classifiers from Samples », *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, ACM, New York, NY, USA, p. 497-503, 2005.

Leite R., Brazdil P., « Active Testing Strategy to Predict the Best Classification Algorithm via Sampling and Metalearning », *Proceedings of the 2010 Conference on ECAI 2010 19th European Conference on Artificial Intelligence*, IOS Press, Amsterdam, The Netherlands, The Netherlands, p. 309-314, 2010.

Lin Y.-S., Jiang J.-Y., Lee S.-J., « A Similarity Measure for Text Classification and Clustering », *IEEE Transactions on Knowledge and Data Engineering*, vol. 99, p. 1, 2013.

Molina M. M., Luna J. M., Romero C., Ventura S., « Meta-learning approach for automatic parameter tuning A case study with educational datasets », *Proceedings of the 5th International Conference on Educational Data Mining, EDM 2012*, p. 180-183, 2012.

Pavón R., Díaz F., Laza R., Luzón V., « Automatic Parameter Tuning with a Bayesian Case-based Reasoning System. A Case of Study », *Expert Syst. Appl.*, vol. 36, n° 2, p. 3407-3420, March, 2009.

Peng Y., Flach P. A., Soares C., Brazdil P., « Improved Dataset Characterisation for Meta-learning », *Proceedings of the 5th International Conference on Discovery Science*, DS '02, Springer-Verlag, London, UK, UK, p. 141-152, 2002.

Pfahringer B., Bensusan H., Giraud-Carrier C. G., « Meta-Learning by Landmarking Various Learning Algorithms », *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 743-750, 2000.

Reif M., Shafait F., Dengel A., « Prediction of Classifier Training Time Including Parameter Optimization », *Proceedings of the 34th Annual German Conference on Advances in Artificial Intelligence*, KI'11, Springer-Verlag, Berlin, Heidelberg, p. 260-271, 2011.

Reif M., Shafait F., Dengel A., « Meta-learning for Evolutionary Parameter Optimization of Classifiers », *Mach. Learn.*, vol. 87, n° 3, p. 357-380, June, 2012.

Rice J. R., « The Algorithm Selection Problem », *Advances in Computers*, vol. 15, p. 65-118, 1976.

Smith-Miles K. A., « Cross-disciplinary Perspectives on Meta-learning for Algorithm Selection », *ACM Comput. Surv.*, vol. 41, n° 1, p. 61-625, January, 2009.

Sun Q., Pfahringer B., « Pairwise meta-rules for better meta-learning-based algorithm ranking », *Machine Learning*, vol. 93, n° 1, p. 141-161, 2013.

Wajeed M., Adilakshmi T., « Different similarity measures for text classification using KNN », *Computer and Communication Technology (ICCCT), 2011 2nd International Conference on*, p. 41-45, Sept, 2011.

Wolpert D. H., Macready W. G., « No free lunch theorems for optimization », *Evolutionary Computation, IEEE Transactions on*, vol. 1, n° 1, p. 67-82, 1997.