Sélection de Caractéristiques pour le Filtrage de Spams

Kamilia MENGHOUR, Labiba SOUICI-MESLATI

Laboratoire LRI, Université Badji Mokhtar,
BP 12, 23000, Annaba, Algérie.
k_menghour@yahoo.fr, souici_labiba@yahoo.fr

RÉSUMÉ. *La sélection des caractéristiques est une étape importante dans les systèmes de classification. Elle vise la réduction du nombre de caractéristiques tout en essayant de préserver ou d'améliorer la performance du classifieur utilisé. Dans cet article, nous proposons une démarche de sélection de caractéristiques, basée sur l'apprentissage automatique, dans le contexte du filtrage de spams qui est considéré comme une tâche de catégorisation de textes. Notre approche consiste à évaluer individuellement chacun des attributs décrivant les messages textuels afin d'ordonner les caractéristiques puis en sélectionner un sous-ensemble suite à une évaluation de performances effectuée en utilisant des classifieurs bayésiens (Naive Bayes) ou de type SVM (Support Vector Machines). Nous avons entrepris une comparaison expérimentale en testant plusieurs combinaisons qui correspondent à des variations des types de classifieurs, des stratégies de sélection (forward/backward) et des méthodes d'évaluation individuelle des attributs, et nous avons obtenu des résultats intéressants. En effet, dans certains cas, nous avons abouti à une réduction significative du nombre de caractéristiques, accompagnée d'une amélioration des performances.*

ABSTRACT. *Feature selection is an important step in classification systems. It aims at reducing the number of features while trying to preserve or improve classifier performance. In this article, we propose a machine learning based feature selection approach, in the context of spam filtering which is considered as a text categorization task. Our approach consists in an individual evaluation of each attribute describing the textual messages in order to sort the features and then select a subset according to a performance evaluation that uses bayesian or SVM classifiers (Naive Bayes or Support Vector Machines respectively). We carried out an experimental comparison by testing several combinations which correspond to variations of classifier types, selection strategies (forward/backward) and individual feature evaluation methods, and we obtained interesting results. Indeed, in some cases, we achieved a significant reduction of the feature number in addition to a performance improvement.*

MOTS-CLÉS : *Sélection de caractéristiques, Filtrage de spams, Classification, Apprentissage automatique.*

KEYWORDS: *Feature selection, Spam filtering, Classification, Machine learning.*

1. Introduction

Le spam est un phénomène mondial et massif. Selon la CNIL (La Commission Nationale de l'Informatique et des Libertés), le spam est défini de la manière suivante : *« Le "spamming" ou "spam" est l'envoi massif, et parfois répété, de courriers électroniques non sollicités, à des personnes avec lesquelles l'expéditeur n'a jamais eu de contact et dont il a capté l'adresse électronique de façon irrégulière. »* (Guillon, 2008). Il existe de nombreuses mesures techniques contre le spam qui peuvent être divisées en deux groupes (Zdziarski, 2005) (Sanz et al, 2008) (Kagström, 2005). Le premier contient les solutions basées sur l'en-tête du message électronique telles que les listes noires, blanches et grises, la vérification de DNS et l'utilisation d'adresse cachée. Le deuxième groupe de solutions contient celles qui sont basées sur le contenu textuel du message telles que le filtrage primitif de contenu, le filtrage heuristique et le filtrage basé sur l'apprentissage automatique.

Dans la littérature il existe de nombreux travaux qui traitent le problème de filtrage de spams en utilisant des méthodes d'apprentissage automatique. Dans (Guzella et al, 2009) les auteurs présentent une revue complète des développements récents dans l'application des algorithmes d'apprentissage automatique pour le filtrage des spams. Le filtrage de spams basé sur le contenu textuel des messages peut être considéré comme un exemple de catégorisation de textes (text categorization) qui consiste en l'attribution de documents textuels à un ensemble de classes prédéfinies (Sanz et al, 2008) (Sebastiani, 2002) (Zhang et al, 2004). Quelle que soit la méthode d'apprentissage utilisée, afin de traiter et tester le filtrage de spam basé sur l'apprentissage automatique, il est nécessaire de construire un large corpus de spams et de messages légitimes. Les e-mails doivent être prétraités pour extraire leurs caractéristiques. Puisque le nombre des caractéristiques dans un corpus peut être très élevé, il est intéressant de sélectionner les caractéristiques qui représentent le mieux les messages avant d'effectuer l'apprentissage du filtre.

La sélection de caractéristiques (feature selection) est un domaine très actif depuis quelques années. Il s'agit de choisir un sous-ensemble minimum de M caractéristiques à partir d'un ensemble original en contenant N ($M \leq N$), de sorte que l'espace de caractéristiques soit réduit de façon optimale selon certains critères d'évaluation (Liu et al, 2005). L'objectif principal est la réduction du nombre de caractéristiques utilisées tout en essayant de maintenir ou d'améliorer les performances de classification du système. La sélection de caractéristiques a été largement étudiée dans plusieurs domaines comme la bioinformatique, la catégorisation de texte, le Data Mining, le traitement d'images…etc (Dash et al, 1997) (Jensen, 2005). Dans (Mendez et al, 2007), les auteurs présentent une analyse détaillée montrant l'influence du changement de la dimension de la représentation d'un message la précision sur certaines techniques classiques de filtrage de spams.

Après avoir étudié le problème de la détection de spams, et constaté le manque de travaux relatifs à la sélection de caractéristiques pour la résolution de ce problème, nous avons entrepris un travail sur l'application de méthodes de sélection de caractéristiques dans le cadre du filtrage de courrier électronique indésirable.

Dans cet article, nous proposons une démarche de sélection de caractéristiques qui consiste à évaluer individuellement chacun des attributs décrivant les messages textuels à l'aide d'une mesure d'évaluation (relief, s2n, gain d'information, information mutuelle, fréquence d'apparition d'un terme dans un document et Chi^2) afin d'ordonner les caractéristiques. Ensuite, un sous-ensemble des caractéristiques ordonnées est évalué par un classifieur de type SVM ou Naïve Bayes. S'il ne donne pas la performance recherchée, nous choisissons un nouveau sous-ensemble et nous réitérons l'apprentissage. Pour construire ce nouveau sous ensemble, nous avons utilisé l'une des deux stratégies forward ou backward. Dans la première, à chaque itération on ajoute une caractéristique, et dans la deuxième on en supprime une.

La suite de cet article est organisée de la manière suivante. La section 2 donne un aperçu des différentes approches de la sélection de caractéristiques. La section 3 est consacrée à la démarche proposée. Nous y introduisons l'approche de sélection de caractéristiques utilisée, nous décrivons brièvement les mesures utilisées pour l'ordonnancement des caractéristiques (feature ranking) ainsi que la mesure utilisée pour l'évaluation des performances de classification. Dans la section 4, nous présentons les performances pendant les phases d'apprentissage et de test. La section 5 est consacrée à une discussion des résultats obtenus. Elle est suivie de la conclusion et de quelques perspectives en section 6.

2. Approches de Sélection de Caractéristiques

Liu et Yu (Liu et al, 2005) ont développé trois dimensions pour catégoriser les méthodes de sélection de caractéristiques. Les stratégies de recherche (complète, séquentielle et aléatoire), les critères d'évaluation (Filtre, Wrapper, et Hybride) et les tâches de data mining (classification ou clustering).

Plusieurs auteurs préfèrent séparer les méthodes de sélection de caractéristiques en deux approches Filtre et Symbiose (Filter et Wrapper) selon leur dépendance ou indépendance par rapport à l'algorithme d'induction (Cakmakov et al, 2002) (Liu et al, 2005). Les méthodes de l'approche Filtre exploitent les propriétés intrinsèques des caractéristiques utilisées, sans référence à une quelconque application. Les méthodes de type Wrapper ou Symbiose définissent la pertinence des caractéristiques par l'intermédiaire d'une prédiction de la performance du système final. Les premiers sont généralement moins coûteux mais aussi moins efficaces que les seconds.

Parmi les méthodes de sélection de caractéristique les plus utilisées actuellement, on trouve les méthodes basées sur l'évaluation individuelle des caractéristiques. Ces procédures de recherche évaluent les caractéristiques individuellement et construisent le sous-ensemble de caractéristiques. Les méthodes connues sous le nom de Feature ranking search (FRS) sont des représentants de la classe des méthodes basées sur l'évaluation individuelle des caractéristiques. Dans ces méthodes, après le calcul de l'utilité des caractéristiques individuelles, on choisit les m meilleures (ou seules) qui donnent une utilité dépassant une certaine valeur seuil.

Si on a déjà consacré des efforts pour obtenir l'utilité des différentes caractéristiques, de manière intuitive, il est raisonnable d'utiliser cette variante simple de construction de sous-ensembles, mais les approches qui sont un peu plus sophistiquées peuvent donner de meilleurs résultats (Cakmakov et al, 2002).

3. Démarche Proposée pour la Sélection de Caractéristiques

Nous proposons une démarche de sélection de variables suivant un processus à deux étapes (voir figure 1). La première étape effectue un filtrage des caractéristiques avec une méthode de type feature ranking (évaluation individuelle des caractéristiques), et la deuxième étape évalue le sous ensemble sélectionné selon l'approche wrapper ou symbiose.

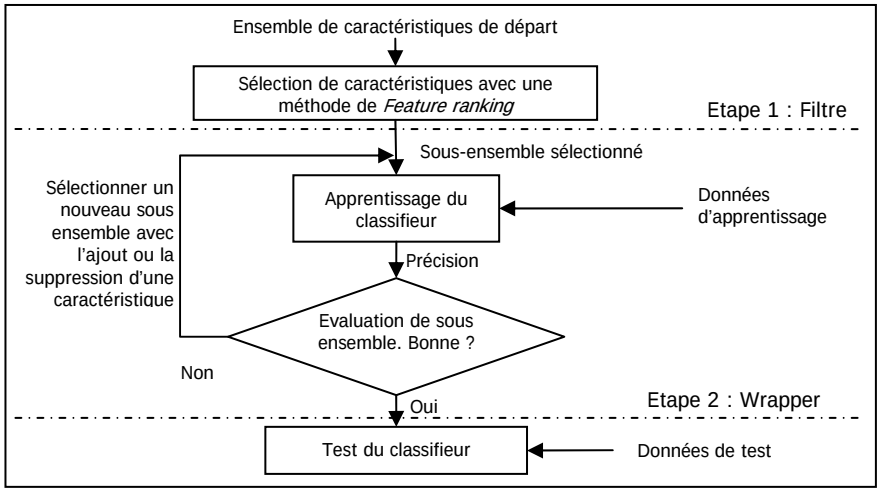


Figure 1. *Démarche proposée pour la sélection de caractéristiques.*

Notre procédure de sélection est une procédure itérative. A chaque itération, nous calculons le poids de chaque caractéristique à l'aide d'une mesure d'évaluation. Ces poids sont utilisés pour ordonner puis sélectionner les f caractéristiques les plus pertinentes parmi toutes les caractéristiques (f étant le nombre des caractéristiques utilisées dans l'itération). Ensuite, cet ensemble sera évalué par un classifieur. Si cet ensemble ne donne pas la solution recherchée, nous choisissons un nouveau sous ensemble et nous réitérons l'algorithme d'apprentissage. Pour générer ce nouveau sous ensemble nous avons utilisé deux stratégies : forward (ascendante) ou backward (descendante). Dans la première, à chaque itération, nous ajoutons la meilleure caractéristique, et dans la deuxième, nous en supprimons la moins bonne. La procédure de sélection s'arrête dans le cas de la stratégie backward si la suppression d'une caractéristique augmente l'erreur de classification. Par contre, dans la stratégie forward la procédure s'arrête si l'erreur d'apprentissage de l'ensemble courant est inférieure à un seuil prédéfini (le seuil utilisé dans notre cas est l'erreur d'apprentissage avant sélection).

3.1. *Méthodes utilisées pour le feature ranking*

Nous avons utilisé les mesures suivantes pour le feature ranking : la fréquence d'apparition d'un terme dans un document DF, l'information mutuelle IM, le gain d'information GI, Relief, signal to noise s2n et Chi^2.

3.1.1. *Document Frequency (DF)*

DF correspond à la fréquence d'apparition des termes dans la collection, c'est le nombre de documents dans lesquels un terme apparaît. On calcule le DF pour chaque terme unique dans le corpus d'apprentissage et on enlève de l'espace de caractéristiques les termes dont le DF est inférieur à un seuil prédéterminé. L'hypothèse de base est que les termes rares sont soit non informatifs pour la prédiction de catégorie, ou non influents sur la performance globale. Dans les deux cas, la suppression des termes rares réduit la dimensionnalité de l'espace de représentation.

3.1.2. *L'information mutuelle (IM)*

L'information mutuelle est l'un des critères les plus utilisés en statistiques. Le critère de l'information mutuelle est calculé comme suit :

$$\mathrm{IM}(w, c) = \log \frac{P(w, c)}{P(w) * P(c)}$$

Avec P (w, c) la proportion de documents dans la collection appartenant à la classe c et où la caractéristique w est présente, P(w) est le DF de la caractéristique w et P(c) est la proportion de documents de classe c.

3.1.3. *Le gain d'information (GI)*

Le gain de l'information mesure la diminution de l'entropie, ou la quantité de connaissance gagnée si une caractéristique est présente ou absente. Le gain d'information (GI) pour un caractéristique w est défini avec par :

$$\mathrm{GI}(w) = \sum_{X \epsilon \{w, \bar{w}\}} \sum_{Y \epsilon \{c_i\}} P(X, Y) \log \left(\frac{P(X, Y)}{P(X) * P(Y)} \right)$$

Avec P (X,Y) la proportion de documents dans la collection appartenant à la classe Y et où la caractéristique X est présente, P(X) est le DF de la caractéristique X et P(Y) : proportion de documents de classe Y dans le corpus.

3.1.4. *La méthode du Chi^2 (χ^2)*

La statistique du χ^2 mesure l'écart à l'indépendance entre une caractéristiques (mot) wk (présente ou absente) et une classe ci (présente ou absente).

	caractéristique w présente	caractéristique w absente	
classe spam présente	a	c	a + c
classe spam absente	b	d	b + d
	a + b	c + d	N = a + b + c + d

Tableau 1 : *Tableau de contingence pour l'absence ou la présence d'un descripteur dans la classe spam.*

Le calcul de Chi^2 nécessite de construire le tableau de contingence (2×2) pour chaque caractéristique w du corpus et pour chaque classe ci (voir tableau 1). Dans notre cas, nous avons seulement 2 classes, donc il suffit de calculer le tableau pour une seule classe.

La statistique du χ^2 peut se mettre sous la forme :

$$\chi^2_{\text{uni}}(\text{w}, \text{spam}) = \frac{\text{N}(\text{ad} - \text{cb})^2}{(\text{a} + \text{c})(\text{b} + \text{d})(\text{a} + \text{b})(\text{c} + \text{d})}$$

Cette formule est calculée pour tous les couples (w, ci). Dans notre cas, il suffit de calculer la valeur de Chi^2 pour la classe spam car $\chi^2(\text{w}, \text{spam}) = \chi^2(\text{w}, \text{légitime})$

Généralement, la valeur discriminante globale de chaque caractéristique (w) est calculée par la mesure: $\chi^2_{\text{max}} = \max_{\text{i}} \chi^2_{\text{uni}}(\text{w}, \text{c}_\text{i})$.

Comme les valeurs de Chi^2 pour les deux classes sont égales, alors la valeur discriminante globale pour une caractéristique (w) est la valeur de cette caractéristique pour la classe spam.

3.1.5. *La méthode signal to noise (s2n)*

Signal-to-noise ratio coefficient (s2n) est une méthode basée sur la corrélation entre les caractéristiques. Cette méthode classe les caractéristiques par le rapport de la valeur absolue de la différence des moyennes des classes sur la moyenne des écart-types des classes. Ce critère est similaire au critère de Fisher et au coefficient de corrélation de Pearson (Dreyfus et al, 2006). La formule de s2n pour une caractéristique w est calculée comme suit:

$$\text{s2n} = \frac{\left|M_{w\,spam} - M_{w\,legitime}\right|}{0.5 * (\sigma_{\text{spam}} + \sigma_{\text{légitime}})}$$

- $\text{M}_{\text{w spam}}$ est la moyenne des valeurs des exemples de la classe spams pour la caractéristique w.
- $\text{M}_{\text{w légitime}}$ est la moyenne des valeurs des exemples de la classe légitime pour la caractéristique w.
- σ_{c} est l'écart-type des valeurs des exemples de la classe c (spams ou légitimes) pour la caractéristique w.

3.1.6. *La méthode Relief*

Relief utilise une méthode statistique pour sélectionner les caractéristiques pertinentes. Il s'agit d'un algorithme basé sur le poids d'une caractéristique inspiré des algorithmes d'apprentissage à base d'exemples (Dash et al, 1997).

L'idée principale de l'algorithme est d'estimer la qualité des caractéristiques selon le fait suivant : « à quel point leurs valeurs distinguent les exemples qui sont proches entre eux ?». Dans ce but, étant donné un exemple aléatoirement choisi X à partir d'un ensemble de données S avec k caractéristiques, Relief recherche l'ensemble de données pour ses deux voisins les plus proches : un de la même classe, appelé nearest hit H, et l'autre d'une classe différente, appelé nearest miss M. Il met à jour l'estimation de la qualité W [Ai] pour toutes les caractéristiques Ai basée sur

les valeurs de la fonction de différence diff () pour X, H et M. m fois, où m est un paramètre défini par l'utilisateur. Pour les exemples, X1, X2, la fonction diff (Ai, X1, X2) calcule la différence entre les valeurs (X1i, X2i) pour la caractéristique Ai:

$$\mathrm{diff}(A_i, x_i, x_i) = \begin{cases} |x_{1i} - x_{2i}| \text{ si } A_i \text{ est numérique} \\ 0 \text{ si } A_i \text{ est nominale et } x_{1i} = x_{2i} \\ 1 \text{ si } A_i \text{est nominale et } x_{1i} \neq x_{2i} \end{cases}$$

La qualité W [Ai] est mise à jour comme suit :

W[Ai] := W[Ai]- diff(Ai,X,H)/m + diff(Ai,X,M)/m;

3.2. *Choix des classifieurs utilisés*

Dans le cadre de nos expérimentations, nous avons choisi d'utiliser deux méthodes d'apprentissage automatique pour la classification : les Support Vector Machines (SVM) et le Naïve Bayes.

Notre choix de ces deux méthodes est basé sur l'étude que nous avons effectuée sur la compétition NIPS 2003 (organisée dans le cadre de la conférence *Neural Information Processing Systems,* sur le thème de la sélection de caractéristiques) et le projet Clopinet (Guyon et al, 2006) et sur les travaux qui ont été menés sur le problème de filtrage de spams. Plusieurs chercheurs concluent dans leurs travaux relatifs au filtrage de spams, à l'aide de classifieurs textuels, que le filtre Bayésien Naive Bayes est la technologie anti-spam la plus efficace. En outre, notre étude des travaux sur la sélection de caractéristiques effectués dans le cadre de la compétition NIPS 2003 et du projet Clopinet (http://clopinet.com), montre que la plupart des gagnants de ces compétitions ont utilisés des classifieurs de type SVM.

D'après la littérature, il n'existe pas de standard pour déterminer les bons paramètres d'un classifieur SVM. Cependant le problème étudié dans notre expérimentation est un problème de filtrage de spams qui peut être considéré comme une tâche de catégorisation de texte. Comme nous nous basons, dans notre travail, sur l'étude que nous avons effectuée sur le projet de NIPS 2003 et du projet Clopinet, nous avons choisi d'utiliser comme paramètres, pour les SVM, les mêmes paramètres que ceux utilisés dans la meilleure entrée pour la base Dexter (ensemble de données de classification des textes), donc nous avons utilisé un noyau polynomial de premier degré.

3.3. *Evaluation des performances de classification*

Pour comparer les résultats issus de ces différents algorithmes, nous allons utiliser l'erreur d'apprentissage BER (Balanced Error Rate) pour évaluer les performances du modèle sachant que, dans notre cas, un exemple positif est un message spam et un exemple négatif est un message légitime.

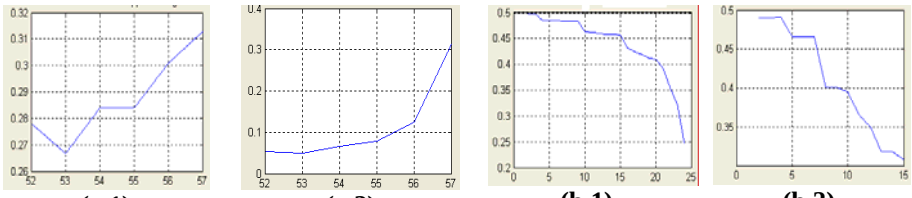
BER: est la moyenne de taux d'erreur de la classe positive et de taux d'erreur de la classe négative = (Erreur des faux positifs + Erreur des faux négatifs) / 2 =

(nombre des exemples positifs mal classés/nombre total des exemples positifs + nombre des exemples négatifs mal classés/nombre total des exemples négatifs) / 2.

4. Résultats obtenus

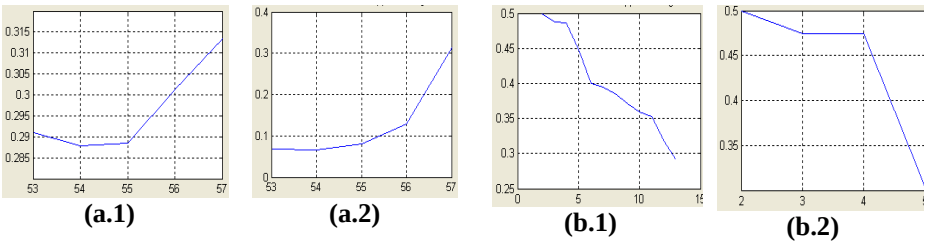
Pour évaluer notre démarche, nous avons utilisé la collection SpamBase disponible dans UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets). Elle contient les informations relatives à 4601 messages, avec 1813 (39,4%) de spam. Cette collection a été prétraitée, et les textes des messages ne sont pas disponibles. Chaque message est décrit en fonction de 57 attributs, le 58ème correspond à la classe du message (spam ou message légitime). Les caractéristiques (55-57) mesurent la longueur des séquences de lettres majuscules consécutives. Nous avons divisé la base SpamBase en 2 ensembles : un ensemble d'apprentissage contenant 3250 exemples (avec 1250 spams), et un ensemble de test contenant 1351 exemples (avec 583 spams).

Les courbes des figures 2 jusqu'à 7 comportent les erreurs d'apprentissage pour les modèles correspondant à chacune des méthodes de sélection proposées dans la section 3.1 avec les deux classifieurs (SVM et Naïve Bayes) et les deux stratégies (Forward et Backward). Dans chacune de ces figures, nous avons utilisé la numérotation suivante : (a.1) la stratégie backward avec le classifieur Naïve Bayes ; (a.2) la stratégie backward avec le classifieur SVM ; (b.1) la stratégie forward avec le classifieur Naïve Bayes ; (b.2) la stratégie forward avec le classifieur SVM.



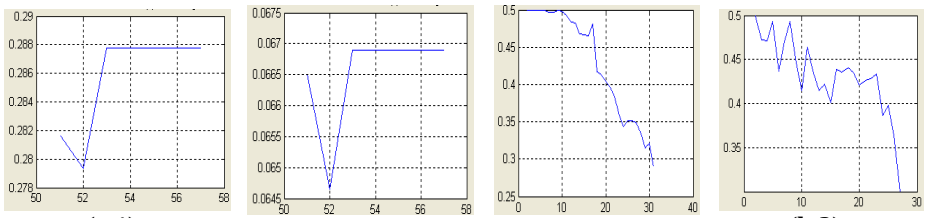
(a.1) **(a.2)** **(b.1)** **(b.2)**

Figure 2. *Variation d'erreur d'apprentissage en fonction du nombre de caractéristiques pour la méthode de sélection DF*



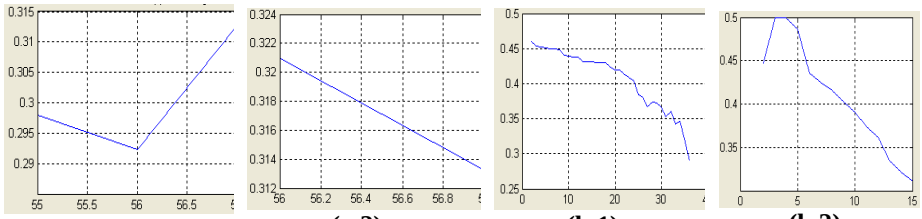
(a.1) **(a.2)** **(b.1)** **(b.2)**

Figure 3. *Variation d'erreur d'apprentissage en fonction du nombre de caractéristiques pour la méthode de sélection IM*



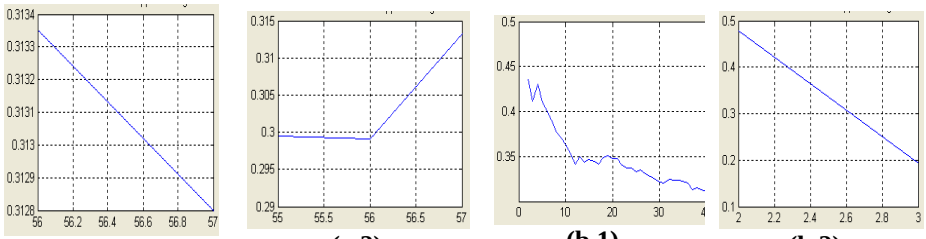
(a.1) **(a.2)** **(b.1)** **(b.2)**

Figure 4. *Variation d'erreur d'apprentissage en fonction du nombre de caractéristiques pour la méthode de sélection GI*



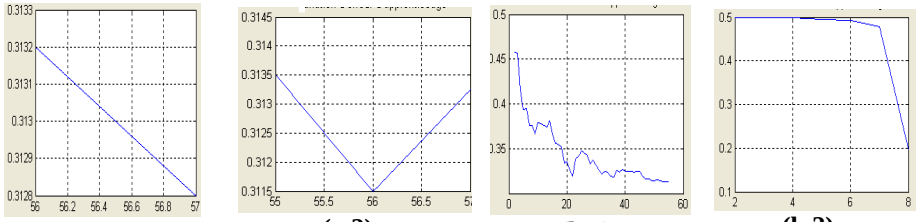
(a.1) **(a.2)** **(b.1)** **(b.2)**

Figure 5. *Variation d'erreur d'apprentissage en fonction du nombre de caractéristiques pour la méthode de sélection Chi²*



(a.1) **(a.2)** **(b.1)** **(b.2)**

Figure 6. *Variation d'erreur d'apprentissage en fonction du nombre de caractéristiques pour la méthode de sélection s2n*



(a.1) **(a.2)** **(b.1)** **(b.2)**

Figure 7. *Variation d'erreur d'apprentissage en fonction du nombre de caractéristiques pour la méthode de sélection Relief*

5. Discussion des résultats

Nous pouvons tout d'abord discuter et comparer les performances de classification sur chaque stratégie de sélection. Les résultats obtenus sur la base d'apprentissage avec toutes les méthodes testées sont résumés dans les tableaux 2 et 3.

	SVM		Naïve Bayes	
	Nb car	Erreur (BRE)	Nb car	Erreur (BRE)
Avant sélection	57	0.3133	57	0.3128
Avec IM	54	0.0669	54	0.2878
Avec IG	52	0.0646	52	0.2793
Avec Chi^2	57	0.3133	56	0.2923
Avec DF	53	0.0499	53	0.2667
Avec S2n	56	0.2991	57	0.3128
Avec Relief	8	0.1963	55	0.3128

Tableau 2. *Résultats obtenus par la stratégie Backward sur la base d'apprentissage*

	SVM		Naïve Bayes	
	Nb car	Erreur (BRE)	Nb car	Erreur (BRE)
Avant sélection	57	0.3133	57	0.3128
Avec IM	5	0.3026	13	0.2913
Avec IG	27	0.3000	31	0.2900
Avec Chi^2	14	0.3219	36	0.2896
Avec DF	15	0.3083	24	0.2467
Avec S2n	3	0.1924	40	0.3105
Avec Relief	8	0.1963	55	0.3128

Tableau 3. *Résultats obtenus par la stratégie Forward sur la base d'apprentissage*

D'après le tableau 2, qui représente les résultats de sélection selon la stratégie Backward, nous constatons que la diminution de nombre de caractéristiques n'est pas très grande avec les deux classifieurs (entre 1 et 5 caractéristiques éliminées seulement). En termes d'erreur d'apprentissage, le classifieur SVM est plus efficace que le classifieur Naïve Bayes et il est capable de diminuer l'erreur de classification d'une façon remarquable (de 0.3 à 0.05). Par contre, avec le Naïve Bayes les méthodes de sélection produisent des taux d'erreur proches (entre 2.6 et 2.9) pour les méthodes information mutuelle, gain d'information, Chi^2 et Document Frequency. Nous constatons aussi une amélioration très faible avec les méthodes S2n et Relief.

D'après le tableau 3, qui représente les résultats de sélection selon la stratégie Forward, nous remarquons, qu'avec le classifieur SVM, la réduction du nombre des caractéristiques est tout à fait significative, particulièrement avec les méthodes de sélection : signal to noise (s2n) qui réduit le nombre à 3 caractéristiques seulement, Relief qui réduit le nombre à 8 caractéristiques, et l'information mutuelle qui réduit le nombre à 5 caractéristiques (avec une amélioration d'erreur d'apprentissage remarquable pour les deux méthodes Relief et S2n de 0.3 à 0.19). Par contre, avec le classifieur Naïve Bayes, la majorité des méthodes ne nous donnent pas des résultats intéressants sauf l'information mutuelle qui réduit le nombre à 13 caractéristiques avec une valeur d'erreur qui avoisine 0.29. Les valeurs de l'erreur de classification obtenue avec 5, 14, 15, et 27 caractéristiques en utilisant le classifieur SVM ne sont pas significativement différentes de celles obtenues sur l'ensemble total des caractéristiques, par contre le nombre de caractéristiques est réduit d'une manière assez importante.

Après la synthèse de nos résultats, différentes remarques peuvent être faites :

- En termes de nombre de caractéristiques sélectionnées, la stratégie forward est plus intéressante par rapport à backward, par contre pour l'erreur de classification la stratégie backward est plus intéressante.
- Quelle que soit la méthode de sélection utilisée dans la stratégie backward, les résultats obtenus avec le classifieur SVM sont meilleurs que ceux du classifieur Naïve Bayes. C'est aussi le cas de manière générale pour la stratégie Forward.
- Les résultats varient selon la méthode de sélection et l'algorithme d'apprentissage. Le meilleur résultat en termes d'erreur est obtenu avec 53 caractéristiques sélectionnées par la méthode DF (Fréquence d'apparition des termes dans la collection) en utilisant un classifieur SVM. Le meilleur résultat, en termes de réduction de nombre de caractéristiques, est obtenu avec 3 caractéristiques sélectionnées par S2n (Signal to noise ratio) en utilisant aussi un classifieur SVM.

	Forward		Backward	
	SVM	Naïve Bayes	SVM	Naïve Bayes
Sans sélection	0.3619	0.2748	0.3619	0.2748
Avec IM	0.3157	0.3239	0.1403	0.2584
Avec IG	0.3009	0.2860	0.1350	0.2551
Avec Chi^2	0.2874	0.2987	0.3619	0.2719
Avec DF	0.2614	0.2719	0.1393	0.2489
Avec S2N	0.3000	0.2746	0.3515	0.2748
Avec Relief	0.3287	0.2739	0.3601	0.2748

Tableau 4. *Taux d'erreur obtenus sur la base de test*

6. Conclusion et Perspectives

La sélection de caractéristique (feature selection) est une étape très importante dans une procédure de classification. Elle consiste à sélectionner un sous ensemble de caractéristiques pertinentes à partir de l'ensemble d'origine. L'objectif étant de réduire le nombre de caractéristiques utilisées, tout en essayant de maintenir ou d'améliorer les performances de la classification.

Dans cet article nous avons proposé une approche de sélection de caractéristiques, pouvant être intégrée dans un processus de construction d'un filtre anti-spams, basé sur les textes des messages considérés, dans le but de réduire le nombre des caractéristiques utilisées et d'améliorer les performances de ce filtre.

Notre démarche de sélection des caractéristiques consiste, dans un premier temps, à effectuer un ordonnancement d'attributs, basé sur l'évaluation individuelle des caractéristiques, et, dans un second temps, à évaluer des sous-ensembles de caractéristiques sélectionnés selon une approche wrapper.

Le travail effectué nous a permis d'explorer le domaine de la sélection de caractéristiques, de faire une comparaison entre plusieurs de ces méthodes dans le cadre de la détection de spams, donnant ainsi quelques indications sur celles qu'il serait intéressant d'utiliser dans une application de ce type.

Bien que les résultats obtenus soient intéressants et encourageants, beaucoup de points sont susceptibles d'être étudiés dans le cadre de travaux futurs, tels que l'utilisation d'autres mesures de sélection dans l'étape d'évaluation individuelle des caractéristiques, l'utilisation d'autres classifieurs dans l'étape de validation de sous ensembles de caractéristiques, l'expérimentation de démarches différentes de celle qui a été proposée dans cet article, et l'utilisation d'autres ensembles d'attributs, en mettant en œuvre les étapes de prétraitement textuels et d'extraction de caractéristiques sur des bases de données de textes de messages électroniques.

7. Bibliographie

Cakmakov D., Bennani Y.*, Feature selection for pattern recognition*, Skopje, Informa, 2002.

Dash M., Liu H., « Feature Selection Methods for Classification », *Intelligent Data Analysis*: An International Journal 1, pp. 131-156, 1997.

Dreyfus G., Guyon I., «Assessment Methods», *Springer Studies in Fuzziness and Soft Computing*, Volume 207, pp. 65-88, 2006.

Guillon P., Etat de l'art du spam, Solutions et Recommandations, Mémoire de Bachelor, Haute école de gestion de Genève, 2008

Guyon I., , Gunn S., Nikravesh M., Zadeh L. A. (Eds.), *Feature Extraction, foundations and applications*, Studies in Fuzziness and Soft Computing, Volume 207, Springer, 2006

Guzella T.S., Caminhas W.M, «A review of machine learning approaches to Spam filtering», *Expert Systems with Applications 36*, pp. 10206-10222, 2009.

Jensen R., Combining rough and fuzzy sets for feature selection, PhD Thesis, University of Edinburgh, 2005.

Kagström J. Improving naive bayesian spam filtering, Master thesis, Mid Sweden University, 2005.

Liu H., Yu L., « Toward Integrating Feature Selection Algorithms for Classification and Clustering », IEEE Trans. on Knowledge and Data Engineering, 17(4), pp. 491-502, 2005

Mendez. J.R., Corzo B., Glez-Peña D., Fdez-Riverola F., Díaz F., «Analyzing the Performance of Spam Filtering Methods When Dimensionality of Input Vector Changes», in P. Perner (Ed.): MLDM 2007, *LNAI 4571*, pp. 364-378, 2007.

Sanz . E.P, Hidalgo J M G, Perez J C C,,«Email spam filtering», in Zelkowitz M. (Ed.), *Advances in computers*, vol.74, pp. 45-114, 2008

Sebastiani F., «Machine Learning in Automated Text Categorization», *ACM Computing Surveys*, Vol. 34, No. 1, March 2002.

Zdziarski J. A., *Ending Spam: Bayesian Content Filtering and the Art of Statistical Language Classification*, No Starch Press, 2005.

Zhang L., Jingbo Z., et Tianshun Y., «An Evaluation of Statistical Spam Filtering Techniques», A*CM Transactions on Asian Language Information Processing*, Vol. 3, No. 4, pp. 1-27, 2004.