Chapitre 5. Evaluation d’un classifieur, expérimentations, et résultats

5.1 Introduction

Comme nous l'avons indiqué auparavant, la décision de classer un document dans une catégorie ou une autre se base sur son contenu et donc elle est *subjective*. Cette subjectivité entraine une difficulté d'évaluation des décisions prises par le classifieur puisqu'on ne dispose pas d'une définition formelle et précise de ce qui rend un document pertinent à une catégorie ou non. Pour cela, les chercheurs font souvent recours à des méthodes empiriques. [Sebastiani, 2002] page 32 confirme le précédent en indiquant « c*omme c'est le cas avec les systèmes de recherche d'information (information retrieval), nous nous basons sur l'expérience pour évaluer un classifieur de textes, plutôt que de procéder analytiquement. La raison en est simple : pour pouvoir évaluer un système analytiquement, on doit posséder une définition formelle du problème à résoudre. Il faudrait pouvoir spécifier à quoi correspondent exactement la rectitude et la complétude. En fait, c'est l'appartenance d'un document à une catégorie, sa pertinence au sein d'une catégorie, qu'il faut définir. Cependant, un caractère très subjectif ressort de ce concept, qui relève aussi de la sémantique d'un texte. Donc, pour l'instant, ce n'est pas formalisable. On se replie alors sur une évaluation empirique des classifieurs* ». Plusieurs mesures d'évaluation empiriques ont été proposées dans la littérature. Nous allons nous contenter de présenter, dans ce chapitre, celles souvent utilisées pas les chercheurs du domaine de la classification automatique des documents.

5.2 Évaluation des résultats du classifieur

Pour évaluer les résultats obtenus pas un classifieur, les documents de l'espace d'apprentissage sont souvent divisés en deux ensembles : le premier est utilisé pour la construction du classifieur tandis que le deuxième est utilisé pour faire le test. Puisqu'on adopte l'approche de classification supervisée on connait à l'avance la catégorie de chaque document. Ainsi, on compare la catégorie prédite avec celle prédéfinie et on calcul un score de performance. Ce calcul peut se faire de diverses façons. Dans ce qui suit, nous allons présenter les méthodes qu'on a utilisé pour mesurer la performance des classifieurs et telles qu'elles étaient présentées par [Sebastiani, 2002]. Il faut noter que ces méthodes sont souvent utilisées dans la littérature.

Pour mieux illustrer les différentes mesures qui vont suivre, on prend pour point de départ la table de contingence illustrée par le Tableau 5.1.

Catégorie c_i		Jugements de l'expert	
		Document appartenant à la catégorie c_i	**Document n'appartenant pas à la catégorie c_i**
Jugements du Classifieur	**Document assigné à la catégorie par le classifieur**	VP_i	FP_i
	Document rejeté de la catégorie par le classifieur	FN_i	VN_i

Tableau 5.1 Table de contingence pour une catégorie

Où,

VP_i (Vrai Positif) est le nombre de documents correctement classés dans la catégorie c_i**,**

FP_i (Faux Positif) est le nombre de documents incorrectement classés dans la catégorie $\boldsymbol{c_i}$**,**

VN_i (Vrai Négatif) est le nombre de documents correctement rejetés,

FN_i (Faux Négatif) est le nombre de documents incorrectement rejetés.

L'efficacité de la catégorisation est normalement mesurée suivant les deux notions classiques de la recherche d'information, notamment, la *précision* (π) et le *rappel* (ρ).

La *précision* (π) par rapport à une catégorie c_i **,** notée par (π_i), est la probabilité conditionnelle qu'un document d_x choisi aléatoirement soit bien classé par le système, autrement dit, $P\big(\breve{\Phi}(d_x, c_i) = 1 \,\big|\, \Phi(d_x, c_i) = 1\big)$.

Le *rappel* (ρ) par rapport à une catégorie c_i , noté par (ρ_i), est la fraction des documents jugés pertinents par le classifieur. Autrement dit, si un document aléatoire

d_x doit être classé dans la catégorie c_i, cette décision est prise. Formellement, on représente le rappel comme $P(\Phi(d_x, c_i) = 1 \mid \breve{\Phi}(d_x, c_i) = 1)$.

Selon [Sebastiani, 2002], « *la précision mesure le degré de la rectitude d'un classifieur tandis que le rappel mesure le degré de sa complétude* ». En utilisant le Tableau 5.1, une estimation de la précision et du rappel par rapport à une catégorie c_i est calculée, comme :

$$\hat{\pi}_i = \frac{VP_i}{VP_i + FP_i}$$

et

$$\hat{\rho}_i = \frac{VP_i}{VP_i + FN_i}$$

Ce qui est recherché, en fait, est une estimation globale et non une estimation pour chaque catégorie. Pour cela, il y a deux principales façons de faire une moyenne pour obtenir un score global : la micro-moyenne et la macro-moyenne.

L'ensemble des catégories $\boldsymbol{C} = \{\boldsymbol{c_1}, \boldsymbol{c_2}, \dots, \boldsymbol{c_{\lvert C\rvert}}\}$		Jugements de l'expert	
		Document appartenant à la catégorie $\boldsymbol{c_i}$	**Document n'appartenant pas à la catégorie $\boldsymbol{c_i}$**
Jugements du Classifieur	**Document assigné à la catégorie par le classifieur**	$VP = \sum_{i=1}^{\lvert C\rvert} VP_i$	$FP = \sum_{i=1}^{\lvert C\rvert} FP_i$
	Document rejeté de la catégorie par le classifieur	$FN = \sum_{i=1}^{\lvert C\rvert} FN_i$	$VN = \sum_{i=1}^{\lvert C\rvert} VN_i$

Tableau 5.2 Table de contingence globale

La micro-moyenne regroupe d'abord les données de chaque catégorie dans une même table de contingence globale et calcule ensuite les scores à partir de celle-ci.

$$\hat{\pi}^{\mu} = \frac{\sum_{i=1}^{|C|} VP_i}{\sum_{i=1}^{|C|}(VP_i + FP_i)}$$

$$\hat{\rho}^{\mu} = \frac{\sum_{i=1}^{|c|} VP_i}{\sum_{i=1}^{|c|}(VP_i + FN_i)}$$

Où μ indique la micro-moyenne.

La macro-moyenne calcule d'abord les scores pour chaque catégorie et fait ensuite une moyenne sur ces scores.

$$\hat{\pi}^{M} = \frac{\sum_{i=1}^{|C|} \hat{\pi}_i}{|C|}$$

$$\hat{\rho}^{M} = \frac{\sum_{i=1}^{M} \hat{\rho}_i}{|C|}$$

Où M indique la macro-moyenne.

La distinction à faire entre ces deux méthodes est que « *la macro-moyenne donne une importance égale à toutes les catégories, tandis que la micro-moyenne donne une importance égale à tous les documents* » [Sebastiani, 1999]. Les avis sont partagés quant à savoir laquelle des deux mesures est préférable. Les tenants de la macro-moyenne affirment que la micro-moyenne est moins représentative parce que les catégories les plus fréquentes ont plus de poids que les autres. Et « *c'est justement cet aspect de la micro-moyenne que ses défenseurs apprécient* » [Sebastiani, 1999]. Ainsi, il est toujours préférable, lors de la présentation des résultats d'une expérience, d'identifier clairement la mesure utilisée ce qui permettra une comparaison plus facile avec d'autres expériences.

Lors de l'évaluation de la performance d'un classifieur, on ne peut tenir compte de la précision ou du rappel séparément. « *On pourrait, éventuellement, mettre en place un système qui va classer tous les documents dans toutes les catégories. Cela nous permettra d'avoir un taux de rappel égal à 100% mais, malheureusement, une très basse précision*» [Sebastiani, 1999]. On voit donc que ce vers quoi il faut tendre un classifieur qui fait le compromis idéal entre ces deux facteurs.

Dans cet ordre d'idées, on peut calculer le « seuil de rentabilité » (en anglais, *«break-even point »)* dénotée par $F1$, c'est-à-dire le point où la précision et le rappel sont égaux. Cette mesure est largement utilisée pour évaluer et comparer la performance des classifieurs. Plus ce point se rapproche de 100%, plus le classifieur est performant à la fois en précision et en rappel. Elle est définie comme :

$$F1 = \frac{2\pi\rho}{\pi + \rho}$$

C'est une fonction qui est maximisée quand la précision et le rappel sont proches. On cherche généralement à l'optimiser lors de l'ajustement du seuil. On peut aussi utiliser une forme plus générale de la mesure $F1$ en ajoutant un paramètre qui pondère l'importance relative des deux critères. « *Il se peut en effet qu'une application particulière nécessite une précision élevée, mais puisse se permettre un rappel un peu moins bon et vice versa* » [Yang, 1999].

En vue de cette variété de critères, un point important à considérer, lorsqu'il s'agit d'évaluer en parallèle différents classifieurs, est la comparabilité des mesures de performance utilisées. Chaque chercheur doit donc se faire un devoir de mentionner précisément la façon dont il a procédé pour évaluer son classifieur. Ainsi, la comparaison des différentes techniques de classification automatique, qui apparaît déjà assez complexe, serait un peu plus facile.

5.3 Expérimentations et résultats

Comme indiqué dans le chapitre 4, nous avons construit notre propre espace d'apprentissage à partir des flux RSS et des sites web Arabes. Pour mener nos expérimentations nous avons utilisé le logiciel Weka[94] comme outil de fouille de données. Pour valider nos résultats, nous avons utilisé la technique de « validation croisée à 10 plis stratifiées[95]».

A l'heure actuelle, comparés à ceux qui rapportent sur la classification automatique des documents écrits en caractère latin, peu de travaux rapportent sur la classification automatique des documents écrits en caractères arabes et aucun d'entre eux n'utilise la technique de « Boosting ». C'est pourquoi nous avons mené une étude d'exploration de cette technique et de son efficacité en vue de la classification automatique des documents arabes. L'apprenti faible qu'on a boosté est l'algorithme des arbres de décision C4.5, un descendant récent de l'algorithme ID3. Nous avons aussi comparé la fidélité et la performance des arbres de décisions dopés contre :

1. C4.5 sans dopage,

[94] http://www.cs.waikato.ac.nz/ml/weka/ [Date de dernière visite : Mai 2010].
[95] cf. §4.8

2. Les machines à vecteurs de support,
3. Les réseaux bayésiens naïfs standards,
4. et les réseaux bayésiens naïfs multinomiaux.

Il faut noter qu'on s'est expérimenté avec les souches de décision dopées (en anglais, *« boosted decision stumps »*) qui sont des arbres de décision très basiques. Malheureusement, elles ont produit des résultats très faibles et médiocres même avec un grand nombre d'itérations de dopage i.e. nous sommes allés jusqu'à 1000 itérations et l'exactitude de boosting n'a pas pu dépasser les 33.42%. En plus, la mesure moyenne de F1 était très faible (0.181) et donc nous avons décidé de ne plus les inclure dans cette étude.

L'unité adoptée pour représenter les attributs est *la racine* (voir le Chapitre 4 pour la justification de ce choix). Le corpus d'apprentissage utilisé lors de cette étude contient 6,825 documents de différents tailles et contenus repartis entre sept catégories comme le démontre le tableau ci-dessous.

Catégorie	**# de Documents**
Droit	889
Economie	960
Médicine	1191
Politique	1020
Religion	1165
Science & Technologie	733
Sports	867
Total	**6825**

Tableau 5.3 Les repartitions des documents dans les catégories du corpus

Pareillement à la méthodologie adoptée par les expérimentations du chapitre 4, nous avons conçu des corpus composé d'un petit nombre d'attributs puis nous avons augmenté ce nombre petit à petit i.e. nous avons commencé avec 161 attributs puis 365 ensuite 520 puis 879 et finalement 1239 attributs. Pour obtenir les sous-ensembles d'attributs nous avons utilisé les deux mesures de réduction de la taille du vocabulaire : le gain d'information et le chi carré. Les figures ci-dessous nous

montrent l'exactitude (qui dénote le pourcentage des documents correctement classés) de chacun des classifieurs basée sur les paramètres indiqués ci-dessus[96].

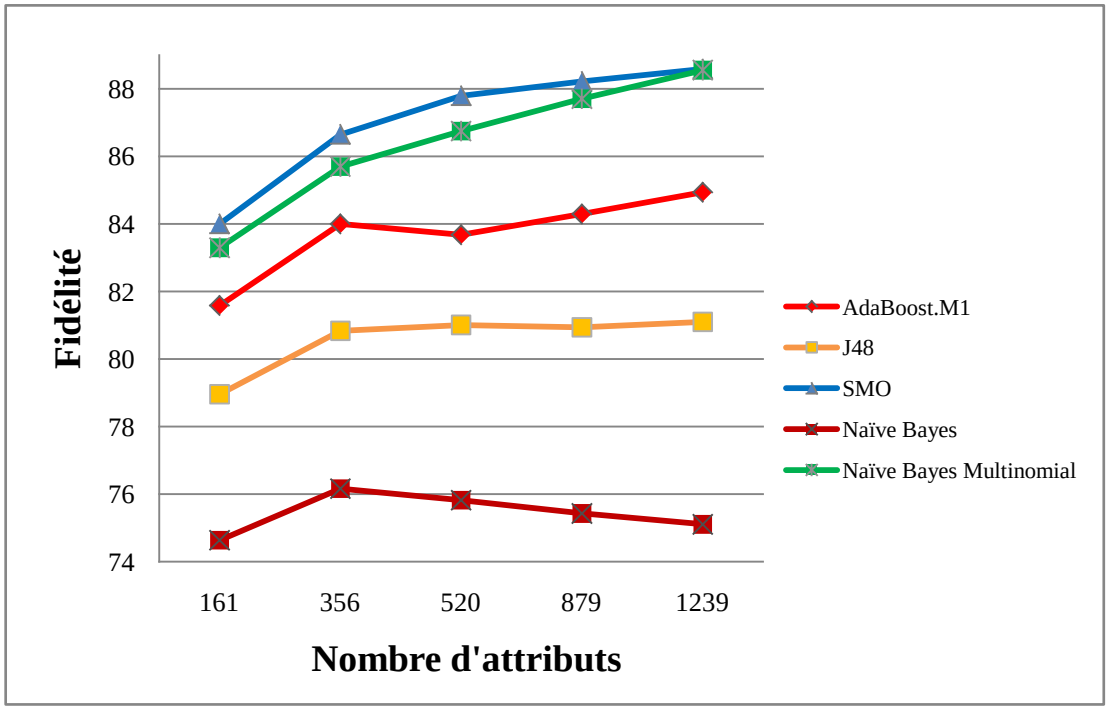

Figure 5.1 Fidélité en utilisant GI

[96] Pour les valeurs des mesures d'évaluation empirique, notamment la précision, le rappel, et F1, voir les figures dans l'annexe page 174.

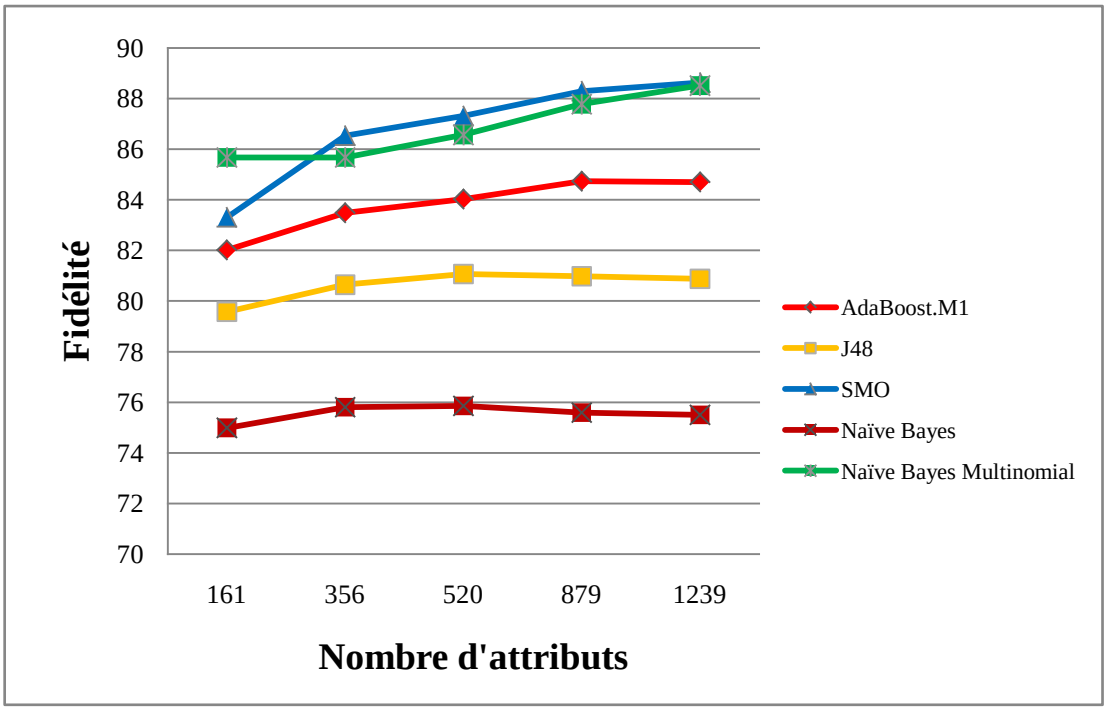

Figure 5.2 Fidélité en utilisant X^2

Il est clair d'après les figures précédentes que les machines à vecteurs de support (SVM) et les réseaux bayésiens naïfs multinomiaux (NBM) surpassent AdaBoost par 3.6% et les autres algorithmes par plus que 8% avec SVM parfois meilleur que les NBM. AdaBoost a bien effectivement boosté et amélioré la performance de C4.5 mais pas suffisamment pour surpasser SVM et NBM. Cela est dû, en principe, au fait que les arbres de décision sont très susceptibles au moindre changement aux données sous-jacentes. En outre, les machines à vecteurs de support sont capables de bien fonctionner dans un espace d'apprentissage ayant une large dimension et de produire de très bons résultats puisqu'ils opèrent indépendamment de la taille de l'espace et se concentre uniquement sur les attributs qui forment les machines à vecteurs de support. Cette propriété rend les machines à vecteurs de support résistibles au sur-apprentissage. De plus, ils sont capables, presque toujours, de trouver l'hyperplan de séparation optimal quant à la classification automatique des documents [Joachims, 1998]. L'algorithme NBM a largement surpassé les réseaux bayésiens naïfs classiques. Cela est peut-être dû au fait que le NBM tient en compte les fréquences et l'ordre des attributs surtout que les racines ont, surement, une fréquence plus élevée que celles des mots ou des lemmes desquels elles sont obtenues.

Nous nous sommes ensuite concentrés sur l'algorithme C4.5 dopé en utilisant le gain d'information comme méthode de réduction de la taille du vocabulaire pour voir si on peut obtenir une meilleure performance que celle obtenu antérieurement. Nous avons augmenté progressivement le nombre d'itérations du dopage. Malheureusement, la performance de C4.5 boosté ne s'est pas améliorée même quand le nombre d'itérations est arrivé à 100. Toutefois, la performance d'AdaBoost ne s'est pas détériorée, comme démontré la Figure 5.3. Cette observation[97] confirme son pouvoir de résister au sur-apprentissage.

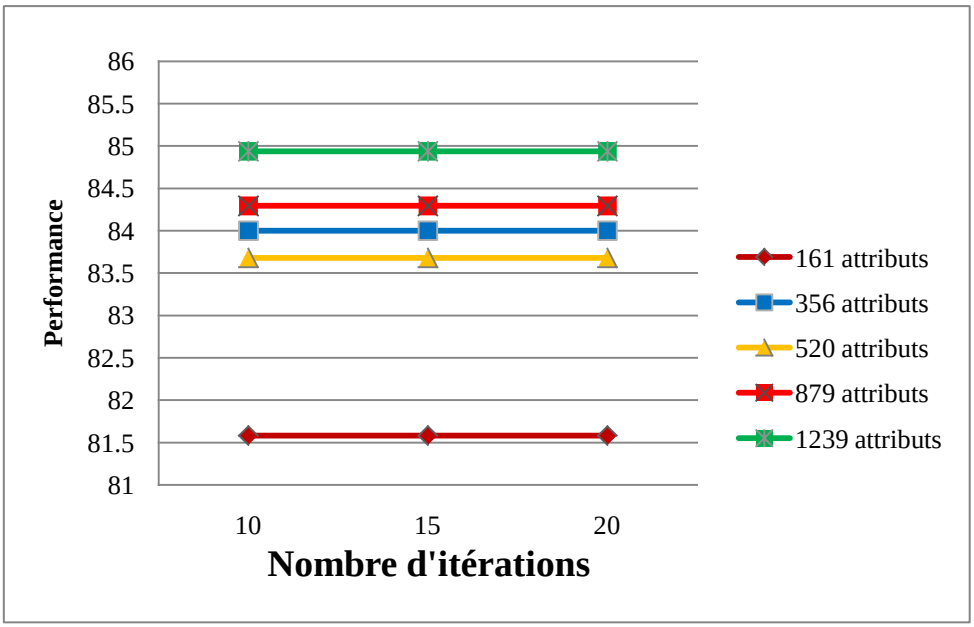

Figure 5.3 Performance de AdaBoost.M1 avec différentes itérations

5.4 Conclusion

Ce chapitre vise à présenter les résultats empiriques de la classification automatique des documents ayant un contenu monolingue en utilisant la technique de Boosting. Nous la comparons avec les algorithmes des machines à vecteurs de support, les réseaux bayésiens naïfs et les arbres de décision. Le SVM ainsi que le NBM restent plus fidèles à la classification automatique des documents que les arbres de décision dopés. Cependant, Boosting pourra donner des meilleurs résultats s'il est utilisé avec

[97] Ce qui confirme un des avantages cités dans le paragraphe §3.7.

un apprenti plus performant que C4.5. Nous avons trouvé que le type d'attribut qui produit les meilleurs résultats est la racine plutôt que les mots dans leurs formes originales ou les lemmes. Nous avons trouvé que le Gain d'Information et le Chi Carré sont les deux techniques de réduction de la taille du vocabulaire qui contribuent mieux à la fidélité de l'apprentissage que les autres techniques.

Dans les chapitres suivants nous allons présenter la classification automatique des documents ayant un contenu multilingue. En particulier, nous allons présenter les avantages et les défis que présente ce genre de documents, les solutions proposées aux problèmes liés à la classification automatique de ces documents ainsi que les expérimentations menées et leurs résultats.