Apport des techniques de Text Mining pour la définition de caractéristiques clefs d'une mammographie

J. Clech*, A. D. Zighed*, A. Brémond**

*Laboratoire E.R.I.C.Université Lumière – Lyon
25 av. Pierre Mendès-France
69 676 BRON cedex – France
jclech@eric.univ-lyon2.fr, zighed@univ-lyon2.fr
**Centre Régional de Lutte Contre le Cancer Léon-Bérard
28, rue Laënnec
69 008 LYON – France
bremond@lyon.fnclcc.fr

Résumé. Dans ce papier, nous exprimons les motivations et la démarche nous ayant conduit à utiliser un corpus de compte-rendus de mammographies élaborés par des radiologues afin de déterminer un ensemble de caractéristiques fréquentes et discriminantes d'une mammographie permettant de prédire la malignité ou la bénignité d'un cas. Pour ce faire, nous avons utilisé des techniques de Data Mining et plus particulièrement de Text Mining pour traiter ce corpus afin d'en extraire dans un premier temps une liste d'une centaine de termes discriminants. Dans un second temps, nous avons mis en œuvre 3 techniques d'apprentissage dans le but de vérifier la pertinence des termes extraits ainsi que de déterminer les termes les plus importants dans les modèles élaborés. Ces derniers ont un taux d'erreur de l'ordre de 21%, mais l'utilisation d'un méta apprentissage définissant la manière de faire coopérer les classifieurs permet d'obtenir un taux d'erreur de 17% ainsi qu'une amélioration significative des taux de sensibilités et spécificités. Nos résultats montrent l'existence d'un riche contenu informationnel au sein des compte-rendus pouvant être pris en compte par nos méthodes nécessitant cependant des travaux complémentaires.

1 Introduction

Les divers programmes de recherche contre le cancer ont apporté leur lot de résultats et de progrès quant aux traitements de ces maladies. Les cancers ne sont pas encore des maladies dont les remèdes sont parfaitement maîtrisés, mais la réussite de leurs applications à des patients est d'autant plus élevée que la détection du cancer est précoce. De ce fait, les campagnes de dépistages de cancers spécifiques, comme le cancer du sein, sont de plus en plus fréquentes. Il en résulte un accroissement important des dossiers à examiner. Dans le cadre de la détection du cancer du sein, les médecins étudient principalement les mammographies des patientes afin de détecter au plus tôt les traces d'un cancer. Par soucis de sécurité, ces mammographies sont examinées par deux radiologues (spécialistes de l'étude de mammographies) sans concertation mutuelle. Dans le cas où leurs avis diffèrent, ils réévaluent ensemble le dossier. En raison de l'accroissement des campagnes de dépistage, et de la nécessité d'une double lecture, les médecins doivent faire face à une surcharge de travail engendrant

potentiellement un diagnostic moins précis dû par exemple à la rareté des cas malins, à la fatigue.

Pour pallier ce problème, des automates se basant sur l'analyse d'images, sont mis en place et se proposent d'effectuer la 2ème lecture de la mammographie en émettant un avis sur la normalité, bénignité ou malignité du cas. Pour ce faire, ils se basent sur un ensemble de caractéristiques préalablement extraites par des techniques de traitements d'images. Ces méthodes offrent des résultats de plus en plus affinés, mais dépendent du choix des caractéristiques à extraire au sein des mammographies. Généralement, ces caractéristiques sont définies à partir de la classification BIRADS de l'American College of Radiology (site ACR) adaptée pour la France par l'Agence Nationale d'Accréditation et d'Evaluation en Santé (site ANAES). En fonction du degré de suspicion, cette classification décrit en 6 catégories les éléments à détecter dans une mammographie. Cependant, n'étant pas son but, cette classification ne fournit pas d'informations sur l'importance relative des différentes caractéristiques proposées ni n'indique celles les plus courantes lors de ces examens. Or, vu la complexité et la difficulté d'extraire ces caractéristiques, il peut paraître préférable de rechercher celles qui apparaissent le plus fréquemment et discriminent le plus les cas normaux des cas malins.

Dans ce papier, nous exprimons les motivations et la démarche nous ayant conduit à utiliser un corpus de compte-rendus (CR) de mammographies indexés par des radiologues afin de déterminer un ensemble de caractéristiques fréquentes et discriminantes. Pour ce faire, nous avons utilisé des techniques de Data Mining et plus particulièrement de Text Mining, mais la mise en œuvre de ces technologies nécessite une approche pluridisciplinaire décomposée en plusieurs étapes. C'est dans ce contexte que s'inscrit cette contribution, conjointement menée avec le Centre Léon Bérard, pôle important de la recherche contre le cancer en France.

Dans un premier temps, section 2, nous donnons les particularités des CR de mammographies et montrons en quoi ils constituent des objets complexes pour le traitement automatique. Dans la section 3, nous abordons la préparation des données. En effet, les CR ne peuvent être pris tels quels et doivent être transformés et mis en forme afin de rendre l'information qu'ils contiennent exploitable par les outils de Data Mining. Cette transformation doit, bien entendu, préserver au maximum le contenu informationnel originel. Dans la section 4, nous décrivons les modélisations effectuées, dans le but de déterminer la qualité informationnelle des caractéristiques extraites. Celles-ci visent à construire des modèles de prédiction. Dans cette expérience, nous nous sommes donc concentrés sur la catégorisation des CR en deux classes : les cas bénins et ceux malins. Dans la section 5, nous établissons un bilan et dressons des perspectives pour l'amélioration et l'exploitation de nos résultats. Nous conclurons enfin en section 6.

2 Particularité des compte-rendus de mammographies

Le CR est un document de travail qui fait partie intégrante du dossier du patient. L'objectif de ce document est de relater une description la plus précise possible des éléments argumentant le diagnostic. Ces éléments sont naturellement ceux présents ou absents des mammographies, mais peuvent également provenir du dossier clinique du patient, permettant de moduler le diagnostic (e.g. âge, actes chirurgicaux précédents). Globalement, les CR de notre corpus sont élaborés sur le schéma suivant : origines des données (e.g. provenance et types des mammographies), description « clinique » des éléments, et une mise en relief de zones éventuellement suspectes. Les CR ne précisent pas toujours de façon explicite si le cas est bénin ou malin. En effet, certains d'entre eux contiennent des conclusions exprimant la

présence ou absence de zones suspectes (e.g. « On ne retient pas d'élément suspect »), d'autres concluent par une nécessité de continuer les investigations (e.g. « On propose une biopsie stéréotaxique »), et d'autres encore ne contiennent pas de conclusion tant l'aspect descriptif est explicite pour le spécialiste.

	Compte-rendus (textes)	***Vocabulaire (mots)***	***Moyenne (occurrences / texte)***
Corpus	734	4223	77 ± 35
Cas bénins	415	2156	80 ± 35
Cas malins	319	2067	74 ± 34

TAB 1 – *Statistiques élémentaires du corpus.*

Dans nos analyses, nous n'allons pas tenir compte des termes intervenant dans ces conclusions. En effet, la catégorisation en tant que telle n'a qu'un apport informationnel minimaliste. Notre objectif est de déterminer quels sont les éléments importants à observer pour l'élaboration d'un diagnostic lors d'une campagne de détection du cancer du sein. Dans ce sens, les modèles permettant la catégorisation seront riches en informations puisqu'ils vont d'une part permettre l'évaluation des termes et d'autre part de déterminer l'impact des différentes caractéristiques sur le diagnostic.

De part leur élaboration, les CR contiennent un vocabulaire très précis et large (4223 mots), et les opérateurs de négation sont nombreux. En outre, les textes sont courts et ont une longueur très variable allant de 30 à 450 mots (Tableau 1).

3 Recherche d'un espace restreint de descripteurs

Puisque l'espace des mots est très creux (1 CR pour 5 mots), une diminution préalable de cet espace est nécessaire. L'objectif de cette étape est donc de déterminer un espace restreint de descripteurs d'une part respectant au mieux l'information originelle, et d'autre part permettant de différencier les 2 classes que nous étudions (bénin, malin).

Pour ce faire, (Luhn 1958) avança l'idée que la fréquence des termes d'un texte apportait une information sur la pertinence de ce dernier et que les termes apparaissant trop fréquemment et trop rarement peuvent être considérés comme faiblement significatifs.

Enfin, dans le cadre de notre problématique, il nous faut prendre en compte également les distributions des termes sur chacune des classes. En effet, si elles sont identiques ou voisines, il est inutile de tenir compte du terme en question dans notre espace de représentation. Pour ce faire, nous pouvons utiliser des méthodes univariées de filtrage comme le OddRatios ou le Chi-2. Cependant, ces familles de méthodes tendent à donner des résultats similaires comme l'explique (Yang et Pedersen 1997).

3.1 Pré-traitement du corpus

Les flexions des termes, i.e. les diverses formes d'un mot, sont légion. Ainsi, la fréquence d'une flexion d'un mot peut être très faible alors que l'ensemble des flexions de ce mot peut être très important. Il est donc intéressant de regrouper sous un même identifiant les différentes formes d'un mot. Pour ce faire, il existe deux méthodes. La première, appelée lemmatisation, consiste à transformer les mots en une forme canonique. Généralement, les formes canoniques utilisées (Lebart et Salem 1994, p37), encore nommées racines linguistiques ou encore lemmes, sont les formes verbales à l'infinitif, les substantifs au singulier, les adjectifs au masculin singulier et les formes élidées à la forme sans élision. La seconde, appelée extraction de pseudo-racines, consiste à extraire une pseudo-racine en prenant en compte les morphologies flexionnelles et dérivationnelles (De Loupy 2001).

L'avantage de la lemmatisation est l'obtention de mots réels et une qualité de regroupement bien supérieur à celle issue d'une phase d'extraction de racines. En effet, la lemmatisation est basée sur une approche linguistique qui permet de lever la plupart des ambiguïtés. Cependant, il est bien évident que le coût calculatoire en est très nettement supérieur comparé à l'approche « mécanique » de l'extraction de racines. Ainsi, les termes *abaissait* et *abaissant* seront regroupés sous le terme abaisser par lemmatisation, et regroupés sous le terme ABAISS par extraction de pseudo-racines. Par contre, le terme *yeux* sera transformé par lemmatisation en ŒIL et sera conservé tel quel par extraction de pseudo-racines.

Pour la réalisation de cette étape, après plusieurs expérimentations, nous avons choisi l'outil d'extraction de pseudo-racines (SNOWBALL), développé par Martin PORTER, pour sa rapidité et la suffisance des résultats regroupant la plupart des formes des mots présents dans notre corpus.

3.2 Sélection des termes et pondération

A partir de cette nouvelle représentation du corpus, nous sélectionnons 50 termes par classe que nous fusionnons ensuite. Pour sélectionner ces 100 termes nous utilisons comme score leurs contributions au Chi-2.

Soient t un terme et c une classe, on définit :

atc : le nombre d'occurrences du terme t appartenant à la classe c.
a.c : le nombre d'occurrences des termes appartenant à la classe c.
at. : le nombre total d'occurrences du terme t toutes classes confondues.
a.. : le nombre total d'occurrences des termes.
Kt, c : la contribution du terme t pour la classe c à la statistique du Chi-2.

Ainsi,
$$K_{t,c} = \frac{\left[a_{tc} - \left(\frac{a_{t.} a_{.c}}{a_{..}} \right) \right]^2}{\left(\frac{a_{t.} a_{.c}}{a_{..}} \right)}$$

Les différents systèmes de pondération de termes sont légions dans la littérature (Aas et Eikvil 1999). La grande majorité est cependant basée sur la fréquence d'un terme pour un

document. L'inconvénient majeur de cette pondération « simple » est de ne pas prendre en compte la fréquence des termes à travers le corpus lui-même.

Nous avons choisi la pondération basée sur le « tf x idf », largement utilisée et palliant sensiblement le défaut précité. Cette pondération corrige la fréquence d'un terme (*term frequency*) par le logarithme de la proportion inverse du nombre de documents du corpus comportant le terme (*inverse document frequency*).

3.3 Listes des termes

Après les étapes de pré-traitement (uniformisation de la casse, regroupement des stemmes, suppression des accents) nous passons d'un espace de 4223 variables à un espace de 795 variables. Le nombre de variables étant du même ordre de grandeur que le nombre de textes, cet espace de représentation est très creux et donc inexploitable tel quel. Pour le réduire davantage, nous sélectionnons alors 50 termes par modalité (bénin, malin) suivant leurs contributions au Chi-2 (Tableau 2).

1- Forme (16)	***2- Texture (18)***	***3- Evolution (14)***	***4- Acte Clinique (14)***
spiculair	irreguli	diminution	chimiotherap
stellair	reguli	appa(rition)	pre-operatoir
asymet(rie)	harmonieu	retraction	depistag
symet(rie)	homogen	conserv	ponction
branche	heterogen	attenu	chirurgical
spicule	hypoechogen	recidif	effectue
canalair	graisseu	augment	realise
punctiform	glandulair	attenuant	ponctionne
mat(rice)	granulair	involution	tentatif
air(e)	infiltrant	ecoul(er)	traite
revet(ement)	flou	surcroit	evalu(er)
architectur	epai	comparabl	cur(e)
foy(é)	densifie	evacu	sond(e)
individualis	densific	deshabite	*arradema(s)*
neoplasm	ruptur		
volum	conjonctivo-glandulair		
	liquid		
	fibro		
5- Entité (13)	***6- Qualificatifs graduels (10)***	***7- Position (10)***	***8- Autre (5)***
kyst	important	contou(r)	*patient*
tumoral	particuli	focalise	*familial*
tumeu	abondant	controlateral	*confront*
lesion	simpl	antero-posterieu	*compte-tenu*
mas(se)	manifest	supero-extern	*cond(ition)*
microcalcific(ation)	retenu	supero-intern	
mastectom	*normal*	local	
mastodyn	*suspect*	isole	
neoplas	*benin*	neo-adjuvant	
adenocarcinom	*malin*	contenu	
fibroadenom			
hamartom			
gynecomast			

TAB 2 - *Liste des 100 termes sélectionnés par contribution au Chi-2.*

Pour une interprétation plus aisée, nous avons regroupé *a posteriori* ces termes en huit catégories (sur le Tableau 2, le nombre de termes appartenant à la catégorie est indiqué entre parenthèses) :

1- Forme : termes regroupant des formes visuelles,
2- Texture : termes qualifiant les différentes textures des formes,
3- Evolution : termes qualifiant l'évolution de zones à risques,
4- Acte Clinique : termes indiquant les actes cliniques effectués,
5- Entité : termes définissant un élément clinique,
6- Qualificatifs graduels : termes modulant le commentaire,
7- Position : termes servant à délimiter une région,
8- Autre : termes n'ayant pu être classés dans les catégories précédentes.

Les termes en italique dans le Tableau 2, n'ont pas été pris en compte pour l'élaboration des modèles. En effet, comme nous l'avons déjà expliqué, les termes nous amenant à une conclusion évidente (normal, suspect, benin et malin) sont ici inintéressants. De plus, le terme arradema(s) correspond en fait au nom d'une campagne de dépistage, et nous est donc totalement inutile. Enfin, les termes affectés à la rubrique 8 sont jugés comme n'apportant aucune information complémentaire.

Ainsi, à l'issue de ces étapes, d'un espace de 4223 variables nous arrivons à un espace bien plus restreint car composé seulement de 91 variables.

4 Modélisations des compte-rendus

Le paradigme de l'apprentissage supervisé, rappelé dans (Zighed et Rakotomomalala 2000), consiste à définir un modèle induit et validé à partir de données *a priori* connues, que nous nommerons par la suite ensemble étiqueté. Par données *a priori* connues, nous entendons les individus (en l'occurrence les CR) dont nous connaissons leur appartenance à une catégorie d'un thème (e.g. le CR x appartient à la catégorie bénin, et le CR y appartient à la catégorie malin). En outre, pour que l'apprentissage soit considéré comme fiable, il faut que cet ensemble étiqueté comporte un nombre suffisant d'individus.

4.1 Méthodes d'apprentissage

Nous avons choisi d'utiliser 3 méthodes usuelles en apprentissage supervisé :

- Arbre d'induction : C4.5,
- Analyse Discriminante,
- Classement par les *k* Plus Proches Voisins (*k*-PPV).

La première, C4.5 (Quinlan 1993), est une méthode permettant la construction puis l'élagage d'arbres d'induction à partir du gain-ratio. Le modèle obtenu est très informatif, car il est composé de règles d'apprentissage de la forme :

Si conjonction de conditions alors conclusion

La seconde, l'analyse discriminante (Fisher 1958), détermine le meilleur hyper-plan (ici une droite) séparateur de nos 2 classes. En outre, les coefficients des variables nous fournissent des informations sur l'intervention des variables dans le modèle.

La troisième, les *k* plus proches voisins (*k*-PPV) (Mitchell 1997), détermine la modalité d'un nouvel individu à partir des *k* individus les plus proches dans l'espace des variables.

Dans notre analyse, nous avons fixé *k* à 10 et utilisé la distance cosine et pondéré les votes proportionnellement à l'inverse de leur distance.

4.2 Méthodes de validation et d'évaluation

Il est important de valider la capacité de généralisation du modèle afin de déterminer la confiance que nous pouvons accorder aux différentes informations ainsi obtenues. Ainsi, pour avoir une mesure plus objective de l'erreur réelle, nous utilisons le principe de la validation croisée (Stone 1974) :

– nous découpons le corpus en *n* parties égales (en conservant la distribution des classes en raison du déséquilibre)

– nous apprenons sur *n-1* parties puis on valide sur la dernière partie, pour les *n* combinaisons possibles

– nous calculons la moyenne des taux d'erreur des *n* validations.

Par ailleurs, nous présentons les taux de sensibilité et de spécificité (dénommés également rappel et précision), afin d'observer et de comparer les qualités des classifieurs.

4.3 Modèles obtenus

Les résultats obtenus à partir des 3 modèles utilisés à la suite d'une 10 validation croisée sont résumés dans les Tableau 3 et Tableau 4. Le premier décrit les taux d'erreur tandis que le second décrit les spécificités et sensibilités.

Au vu de ces tableaux, le résultat le plus évident est l'équivalence des modèles du point de vue de ces critères. De plus, il apparaît une plus grande facilité à reconnaître les cas bénins des cas malins : 10 points de différence pour la sensibilité.

Enfin, l'étude de l'impact des variables dans les modèles de l'analyse discriminante et celui de C4.5 nous indique que l'ensemble des variables ont un impact quasi équivalent.

Modèle	Taux d'erreur	Ecart-type
AD	20%	5%
C4.5	21%	4%
10-PPV	21%	5%

TAB 3 - *Taux d'erreur en validation croisée des 3 modèles.*

	Sensibilité		Spécificité	
Modèle	**Bénin**	**Malin**	**Bénin**	**Malin**
AD	87%	70%	79%	81%
C4.5	83%	74%	81%	77%
10-PPV	83%	73%	80%	77%

TAB 4 - *Taux de sensibilité et spécificité des 3 modèles.*

5 Discussion

Les termes sélectionnés sont intéressants du point de vue sémantique et qualitatif. En effet, ils déterminent un ensemble de caractéristiques discriminantes. Cependant, l'étude de l'impact des variables n'apportant aucun renseignement sur la prévalence de l'une ou l'autre d'entre elles, et ajoutée au fait de la plus grande difficulté de retrouver les cas malins, tendent à nous faire dire qu'il existe un nombre plus élevé de formes pour les cas malins que pour les cas bénins.

De plus, les règles produites par l'arbre d'induction sont à interpréter avec précaution. En effet, il pourrait être tentant d'interpréter au sein de la règle "SI mas(se) ET spicule ALORS MALIN" une association sémantique entre les 2 termes en question. Néanmoins, rien ne nous permet de l'affirmer puisque certes ces 2 termes apparaissent dans le même compte-rendu mais nous n'avons aucune information quant à leur position relative hormis par une étude *a posteriori*. De ce fait, il nous semble nécessaire que les descripteurs utilisés par les modèles doivent intégrer l'information concernant les contextes des termes utilisés.

Les modèles utilisés reflètent des résultats similaires. Cependant, en raison du grand nombre de formes supposés et des 3 paradigmes utilisés, nous nous interrogeons sur la similarité de nos 3 classifieurs. Pour ce faire, nous avons comparé les erreurs (les mauvais classements) de chacun d'entre eux. Nous avons observé que le taux d'erreur communes de 2 classifieurs n'excédait pas les 40%.

Ainsi, nous avons cherché à améliorer nos classifieurs en effectuant un apprentissage sur les résultats de nos 3 modèles. L'idée est de travailler sur leurs erreurs. Au lieu d'utiliser des techniques comme le *bagging* (Breiman 1996) ou le *boosting* (Freund et Schapire 1995), il s'agit plutôt d'étudier la manière dont les modèles pourraient alors coopérer. L'approche que nous préconisons est de construire un méta modèle à partir des prédictions de chacun des 3 modèles. Ce méta modèle tenterait de trouver la meilleur façon d'utiliser les prédicteurs pour améliorer la prédiction. Nous avons alors construit un tableau dans lequel chaque colonne représente le résultat d'un prédicteur pour chaque individu. Nous y avons adjoint la classe à identifier. Nous avons considéré que chaque modèle est un attribut prédictif et nous avons construit un méta modèle en utilisant des algorithmes d'apprentissage classiques comme les graphes d'induction, l'analyse discriminante, etc.

Afin d'obtenir une mesure objective de ce méta apprentissage, nous l'avons évalué dans le cadre d'une 3 validation croisée. De par nos expériences, le meilleur méta modèle est l'analyse discriminante dont les résultats sont décrits en Tableau 5 et Tableau 6.

Modèle	Taux d'erreur	Ecart-type
AD	17%	1%

TAB 5 - *Taux d'erreur en validation croisée du méta modèle.*

	Sensibilité		Spécificité	
Modèle	**Bénin**	**Malin**	**Bénin**	**Malin**
AD	87%	79%	84%	82%

TAB 6 - *Taux de sensibilité et spécificité du méta modèle.*

En comparant ces résultats à ceux obtenus par les simples classifieurs , nous notons une légère augmentation du taux de succès (de 3 à 4 points). Par ailleurs, nous constatons que les sensibilités et spécificités sont supérieurs à toutes les précédentes valeurs.

6 Conclusions et perspectives

Dans ce papier, nous avons posé la problématique de la détection de caractéristiques clefs au sein d'une mammographie. Nous avons montré l'apport potentiel des techniques de Text Mining dans ce cadre. Les expériences menées et décrites dans cet article montrent la possibilité d'exploitation du contenu informationnel des compte-rendus.

Cependant, nous avons mis en exergue certains problèmes liés au choix du type de descripteurs (de simples termes). Trois pistes (non exclusives) nous semblent intéressantes à suivre : utiliser ou compléter des ontologies existantes dans ce domaine, déterminer des descripteurs capables de prendre en compte d'une part les opérateurs de négations et d'autre part les contextes des différents termes, utiliser et comparer d'autres méthodes de sélection de descripteurs comme par exemple RELIEF (Kira et Rendell 1992).

De plus, les résultats de sensibilité et spécificité pour les différents classifieurs montrent qu'il n'est pas équivalent de prédire un cas bénin d'un cas malin. Il nous paraît intéressant d'évaluer l'apport de l'utilisation de fonctions de coûts asymétrique lors de l'élaboration des classifieurs.

Enfin, nous avons montré que l'utilisation d'un méta classifieur peut apporter de la connaissance supplémentaire. Il nous paraît intéressant d'explorer davantage cette voie en la comparant notamment aux techniques existantes agrégeant les règles de différents classifieurs comme le *bagging* ou le *boosting*. Dans ce sens, nous sommes en train de mener des analyses comparatives qui feront l'objet d'une publication ultérieure.

7 Remerciements

Nous tenons à remercier le Centre Léon Bérard – Lyon pour nous avoir fourni le corpus de compte-rendus mais également pour le temps qui a pu nous être consacré.

Références

ANAES http://www.anaes.fr
ACR http://www.acr.org
SNOWBALL http://snowball.tartarus.org

Aas K., Eikvil L. (1999), Text categorisation: A survey. Technical report, Norwegian Computing Center, June 1999.

Breiman L. (1996), Bagging predictors, Machine Learning, vol 24, pp. 123-140, 1996

Fisher W.D., On grouping for maximum of homogeneity. Jour. Ann. Statis. Assoc., p. 789-798, 1958.

Freund Y., Schapire R.E. (1995), A decision theoretic generalization of online learning and an application to boosting. Proceedings of the 2nd European Conference on Computational Learning Theory, Springer Verlag, pp. 137-140, 1995

Kira, K., Rendell. L.A. (1992), A Practical Approach to Feature Selection, 9th International Workshop on Machine Intelligence. Aberdeen, Scotland : Morgan-Kaufman, 1992.

Lebart L., Salem A. (1994), Statistique Textuelle, Paris, Editions Dunod, 1994.
De Loupy C. (2001), L'apport de connaissances linguistiques en recherche documentaire, TALN 2001, Tours, juillet 2001.
Luhn H.P. (1958), The automatic creation of literature abstracts, IBM Journal of Research and Development, 2, 159-165, 1958.
Mitchell T.M. (1997), Machine Learning. Computer Science, McGraw-Hill, New York, 1997.
Quinlan J.R. (1993), C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA, 1993.
Stone M. (1974), Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society. B 36, pp. 111-147, 1974.
Yang Y., Pedersen J.O. (1997), A comparative Study on Feature Selection in Text Categorization, Proc. of the 14th International Conference on Machine Learning,
pp. 412 420, 1997
Zighed D.A., Rakotomomalala R. (2000), Graphes d'induction : Apprentissage et Data Mining. Paris, Editions Hermès, 2000.

Summary

ABSTRACT. In this paper, we define the motivations to use a corpus of mammography reports written by radiologists. From those reports, we select frequent and discriminant characteristics in the aim to use them in prediction of the malignity or benignity of a case. To do that, we apply text mining methods. Then, we define 3 different machine learning methods in order to verify the selected term pertinence and to define the most relevant terms. Their error rates are near 21%, and the using and a meta learning based on the classifiers gives an error rate of 17% and improves the sensibility and specificity rates. Our results show the large information of reports could be taking into account by our methods but require some additional works.