Chapitre 5 Applications sur des données complexes

Résumé. Dans ce chapitre, nous décrivons des applications liées à l'exploitation de données complexes. Nous décrivons le traitement de données multivariées dans le cadre de données alimentaires. Nous avons à cette occasion réalisé une extension gérant l'aspect multivarié au sein de la méthode RECPAM. Nous abordons ensuite des applications liées aux données textuelles. Nous décrivons plus particulièrement notre démarche face à l'analyse de deux corpus de documents textuels atypiques : l'analyse de curriculum vitæ puis l'analyse de compte-rendus médicaux. De plus, nous montrons qu'il est possible de traiter des documents en langues diverses malgré des transformations impliquant une perte du contenu informationnel. Enfin, nous décrivons une méthode permettant le re-étiquetage des catégories d'appartenance des documents, basée sur la notion de contexte et mise en œuvre sur le corpus Reuters Mode Apte.

Mots clefs : Applications, données multivariées, données textuelles, re-étiquetage, graphe de voisinage.

Chapitre 5 – Applications

5.1 Introduction

Dans ce chapitre, nous proposons de présenter une série d'exemples concrets répondant à une problématique de données complexes. La première application, section 5.2, se situe à la marge du traitement de ce type de données puisqu'elle traite ici de données multivariées. La méthode RECPAM permet de répondre à un traitement de données complexes notamment en prenant en compte une interaction multi-niveaux entre les divers descripteurs, mais également en permettant une étude des sous-populations obtenues aux feuilles du graphe. En ce sens, et bien qu'ici elle soit présentée dans une version simplifiée, il nous semble intéressant d'exposer cette application publiée dans (CIAMPI *et al.*, 2000).

Les autres applications sont liées à la problématique de catégorisation de données textuelles. Dans les sections 5.3 et 5.4, nous décrivons des tentatives d'exploiter, d'adapter les outils d'analyse textuelle pour des besoins qualifiés de « métiers » : le secteur du recrutement pour l'analyse des curriculum vitæ et le secteur médical pour l'exploitation des compte-rendus médicaux. Ces deux applications ont été respectivement publiées dans (CLECH et ZIGHED, 2003) et (CLECH *et al.*, 2003b). Enfin, les deux dernières applications décrites en sections 5.5 et 5.6 permettent d'étudier des extensions à la simple catégorisation de documents en intégrant un contexte plus large à sa mise en œuvre : respectivement le traitement des documents multilingues publié dans (JALAM *et al.*, 2004) et celui de documents incorrectement étiquetés publié dans (CLECH et ZIGHED, 2004).

5.2 Traitement de données alimentaires

L'hypothèse qu'un régime alimentaire particulier apporte des bienfaits ou des méfaits sur la santé est déjà connue depuis de nombreuses années, notamment à travers le « régime crétois » constitué d'une alimentation riche en fibres, vitamines et minéraux et d'un peu de vin qui permet de prévenir de l'apparition des maladies cardio-vasculaires. Dans le cadre de la prévention des maladies cancéreuses, une vaste étude épidémiologique à échelle européenne, appelée EPIC, est menée par le Centre International de Recherche contre le Cancer. Cette étude a été lancée en 1990 et inclut 23 centres répartis à travers 10 pays européens. Les données alimentaires de cette étude représentent une population d'environ 500 000 femmes provenant de ces différents centres. Une partie de cette étude a pour objectif d'identifier les régimes alimentaires et de mettre en évidence leurs effets en tant que facteurs de risques liés au cancer (RIBOLI et KAAKS, 1997). Ce type d'analyse doit prendre en compte les interactions potentielles entre les variables aussi bien exogènes qu'endogènes.

Dans un premier temps, nous allons décrire la spécificité de ces données. Puis, nous allons présenter les méthodes que nous avons définies afin d'être en mesure de modéliser de telles données. Enfin, nous appliquerons ces méthodes sur un sous-ensemble de données que nous commenterons.

5.2.1 Caractéristiques des données alimentaires

L'étude EPIC dans sa totalité porte sur 500 000 femmes dispersées dans plusieurs centres à travers différents pays. Ici, nous allons nous intéresser à un ensemble bien plus réduit en n'étudiant que le centre « Ile-de-France » de l'étude, constitué par 1 201 femmes. Les données que nous étudions renseignent sur l'alimentation des sujets durant 24h. Ces données décrivent d'une part la quantité d'aliments absorbés en grammes, et d'autre part l'énergie calorique des nutriments, exprimée en Kilo calories, apportée par ces aliments. Les aliments sont subdivisés en 16 groupes (e.g. Viande, Fruit, Légume, Bouillon) et les nutriments en 4 catégories (Glucides, Lipides, Protéines et Alcool).

Selon les nutritionnistes, étudier un nutriment seul n'a pas de sens particulier tant celui-ci est révélé par l'étude des nutriments dans leur globalité. Nous comprenons que ces variables sont dépendantes, du moins au sens de l'interprétation et de la comparaison entre deux individus. C'est pourquoi nous considérons de telles données comme étant complexes.

5.2.2 La méthode RECPAM

Nous décrivons ici, les principes fondamentaux de la méthode RECPAM ; pour plus de détails voir (CIAMPI, 1991). RECPAM est une méthode arborescente sur laquelle est appliquée une fusion tardive. Elle se déroule en trois étapes, chacune conditionnellement à la matrice de données exogènes D composée de p descripteurs et de la matrice endogène E composée de m variables.

La première étape est celle de la croissance de l'arbre. Elle consiste comme pour CART (BREIMAN *et al.*, 1984) en une construction récursive d'un arbre binaire, maximisant le gain d'information à chaque nœud, jusqu'à ce que les effectifs des feuilles ne dépassent pas un seuil minimal fixé par l'utilisateur. Ici, le gain d'information est défini comme la réduction de l'inertie comme discuté plus loin. Ainsi, à la fin du processus, la structure d'un arbre, composée de nœuds et de feuilles, est obtenue. Dans le vocabulaire de RECPAM, cet arbre est appelé le *grand arbre*.

La seconde étape est celle de l'élagage du grand arbre. La séquence d'élagage est obtenue en minimisant la perte d'information engendrée par la suppression d'une branche de l'arbre. Ainsi, cette opéra-

tion définit un arbre plus petit appelé l'*arbre véritable*. Les branches élaguées du grand arbre sont considérées comme étant virtuelles. Ces éléments virtuels sont utilisés dans la méthode générale de RECPAM pour prendre en compte des interactions subtiles entre les variables endogènes et des variables dites virtuelles. Dans le cadre de cette application, nous ne discuterons pas ce point.

La troisième étape est celle de la fusion tardive. Elle opère sur des partitions, dont la première est formée par les feuilles de l'arbre véritable, et produit dès lors un graphe d'induction. Elle construit de façon récursive une super-partition à partir de la partition de l'étape précédente en fusionnant les deux ensembles dont leur fusion minimise la perte d'information provoquée. L'objectif de cette étape est de simplifier la prédiction et d'accroître la généralisation du modèle.

Ces trois étapes sont certes sous-optimales puisqu'elles suivent un processus glouton, mais cette heuristique permet en pratique d'approcher l'optimum. La principale caractéristique de cette approche est de présenter naturellement la capacité de traiter des données de structure plus complexe moyennant un modèle statistique. En effet, outre la structure arborescente que RECPAM élabore, ce dernier permet la prédiction des paramètres statistiques $\theta = \left[\theta_1, \cdots, \theta_p\right]$ modélisant D : RECPAM peut être vu comme une généralisation des arbres de régression à des modèles statistiques plus larges comme par exemple le modèle de survie (COX), le modèle exponentiel, … et dont les trois étapes du processus de l'arbre sont conditionnées à l'écart de ces données par rapport au modèle statistique utilisé.

Dans la définition la plus restreinte de RECPAM, se rapprochant de CART, l'élaboration du classifieur permettant la construction des paramètres prédictifs est définie en fonction des positions des individus au niveau des G classes fusionnées (à l'issue de la 3ème étape de RECPAM). Pour chacune de ces classes, une constante γ est évaluée (5.2.2) minimisant l'écart au modèle théorique utilisé (5.2.1).

$$\theta_j(d) = \sum_{g=1}^{G} \gamma_g^j I_g(d) \quad \forall j \in [1, p] \tag{5.2.1}$$

$$\widehat{\theta_j}(d) = \sum_{g=1}^{G} \hat{\gamma}_g^j I_g(d) \quad \forall j \in [1, p] \tag{5.2.2}$$

Le modèle général de RECPAM propose d'affiner ce modèle en ajoutant la contribution de certaines variables exogènes. En outre, ces constantes et variables agissent sur trois niveaux différents (du plus global au plus local) en s'appuyant sur la structure arborescente et notamment sur les éléments vir-

tuels. Nous ne détaillerons pas davantage la description de cette méthode car sortant du cadre de cette application.

5.2.2.1.1 Le cas multivarié

Nous nous plaçons dans le cadre où l'endogène *E* est multivariée et assumons alors l'hypothèse de distribution multinormale des données *D* permettant de considérer que le vecteur moyenne et la matrice de variance-covariance de *E* suffisent à décrire les données. Dans ce cas, les paramètres du modèle statistique sont $\theta = \left(\mu_1, \cdots, \mu_p, \mathrm{Cov}\right) = \left(\mu, \mathrm{Cov}\right)$. Le but est de rechercher la structure prédictive de RECPAM la plus adéquate représentant les relations entre *D* et *E*. Notre approche permet de prendre explicitement en compte des associations entre les diverses variables, en contraste avec les approches consistant à construire un arbre pour chacune des variables de *E*.

Pour ce faire, nous nous limitons au cas où nous supposons que la constante est la même pour l'ensemble des composantes de μ et de Cov, bien qu'elle puisse être différente.

Déviance. Supposons que la partition triviale (l'ensemble des données, ou encore la population de la racine de l'arbre) soit en adéquation avec la représentation des données. Alors, pour un couple donné $\left(\mu, \mathrm{Cov}\right)$, une définition naturelle est fonction de la somme des distances de Malahanobis de chaque vecteur moyenne (5.2.3). En supposant une normalité multivariée de ces données, alors une définition encore plus naturelle est moins deux fois le log vraisemblance (5.2.4).

$$dev(D, E, \theta) = \sum_{i=1}^{\mathrm{n}} \left(y^i - \mu\right)' \mathrm{Cov}^{-1} \left(y^i - \mu\right) \tag{5.2.3}$$

$$dev(D, E, \theta) = p \log |\mathrm{Cov}| + \sum_{i=1}^{\mathrm{n}} \left(y^i - \mu\right)' \mathrm{Cov}^{-1} \left(y^i - \mu\right) \tag{5.2.4}$$

Gain d'information. Nous avons vu que la construction de l'arbre se faisait en minimisant la déviance lors du passage de la structure d'arbre T_0 vers la structure *T*. Dès lors, il est naturel de définir le gain d'information comme la différence de ces deux déviances (5.2.5) :

$$IG\left(T : T_0 \mid D, E\right) = dev\left(D, E, \hat{\theta}_{T_0}\right) - dev\left(D, E, \hat{\theta}_T\right) \tag{5.2.5}$$

Estimation des paramètres. Dépendamment des hypothèses spécifiques aux données, plusieurs estimations peuvent être faites. Ici, nous ne présenterons que celles qui nous intéressent. Nous supposons

que la variance-covariance varie, identiquement aux vecteurs moyennes, au niveau des classes fusionnées. Alors, μ est défini par l'expression (5.2.6) alors que Cov est défini par l'expression (5.2.7).

$$\hat{\mu}(z) = \sum_{g=1}^{G} \hat{\mu}_g I_g(z) \tag{5.2.6}$$

$$\hat{\Sigma}(z) = \sum_{g=1}^{G} \hat{\Sigma}_g I_g(z) \text{ avec } \hat{\Sigma}_g = \frac{1}{n_g} \sum_{i \in g} \left(y^i - \hat{\mu}_g\left(\mathrm{z}^{\mathrm{i}}\right) \right)' \left(y^i - \hat{\mu}_g\left(\mathrm{z}^{\mathrm{i}}\right) \right) \tag{5.2.7}$$

5.2.2.1.2 Le cas non supervisé

Dans la sous-section précédente, nous avons vu comment traiter le cas d'un jeu de données comportant une endogène multivariée. Cette approche supervisée fait la distinction naturelle entre les endogènes et les exogènes, alors que dans le cadre non supervisé, nous ne disposons que des exogènes. Nous proposons ici d'adapter RECPAM à ce cadre.

Nous nous intéressons à la recherche d'une structure homogène et des classes distinctes d'individus au sein de ces données. En outre, nous supposons que ces classes doivent être définies à l'aide des variables exogènes, en d'autres termes, nous cherchons à obtenir une classification conceptuelle des données.

Notre approche, que nous nommons Supervision Factorielle, repose sur une proposition plus ancienne de CHAVENT *et al.* (1999). L'idée est de se replacer dans un cadre supervisé à partir des premières composantes principales. Dès lors, nous retrouvons une endogène multivariée, et nous pouvons appliquer la méthode précédemment présentée. Il est à noter que la recherche de telles classes signifie rechercher des classes dans un sous-espace où l'information « pertinente » dispersée à travers les données originales est synthétisée à travers les composantes principales.

5.2.3 Application sur les données alimentaires

Dans une première analyse, essentiellement à titre illustratif, nous avons élaboré un modèle pour expliquer l'énergie des nutriments à partir des 16 catégories alimentaires. L'arbre de la Figure 5-1 a été obtenu par l'approche multivariée décrite ci-dessus. Nous avons obtenu un grand arbre comprenant 5 feuilles, qui a été élagué en un arbre véritable constitué de 3 feuilles. Aucune fusion n'a été effectuée. Deux groupes alimentaires définissent la structure : l'Alcool et la Viande.

- Pour la feuille 1, nous obtenons le profil d'énergie alimentaire suivant (les écart-types sont indiqués entre parenthèses) : 201,101 (76,142) Kcal pour les glucides, 77,047 (33,864) Kcal pour les lipides, 75,139 (25,261) Kcal pour les protéines et 0,006 (0,130) Kcal pour l'alcool ;
- Pour la feuille 2 le profil est : 208,681 (84,166) Kcal pour les glucides, 84,995 (38,957) Kcal pour les lipides, 73,494 (23,879) Kcal pour les protéines et 17,681 (13,701) Kcal pour l'alcool ;
- Pour la feuille 3, nous avons 201,044 (81,174) Kcal pour les glucides, 104,155 (43,092) Kcal pour les lipides, 99,647 (30,547) Kcal pour les protéines et 23,885 (21,198) Kcal pour l'alcool.

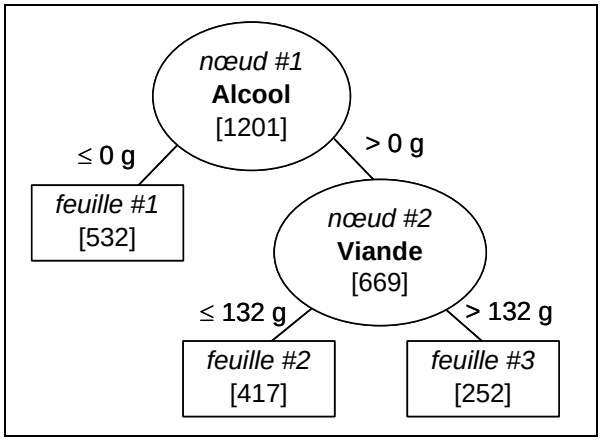


Figure 5-1 Graphe de prédiction multivariée

Les vérifications de nos suppositions sont en cours, mais nous pensons que l'obtention d'un si petit arbre est probablement du à de très faibles variations de régimes alimentaires au sein de la population du centre d' « Ile-de-France ».

La Figure 5-2 a été obtenue en appliquant la supervision factorielle comme décrite plus haut. Préliminairement, une analyse en composantes principales sur les 16 catégories alimentaires a été réalisée. Les 5 premières composantes expliquent plus de 80% de la dispersion. Ces composantes ont été utilisées en tant que variables endogènes lors de la construction du graphe à l'aide de RECPAM. Le grand arbre produit est composé de 7 feuilles et a été élagué à 3 feuilles, formant ainsi l'arbre véritable. L'étape de fusion ramène finalement le graphe à 2 classes. La première est constituée de 271 individus caractérisés par l'absence de consommation de « soupes et bouillons » ni d' « alcool ». La seconde est composée des 830 individus restant consommant au moins l'une ou l'autre des catégories alimentaires citées précédemment.

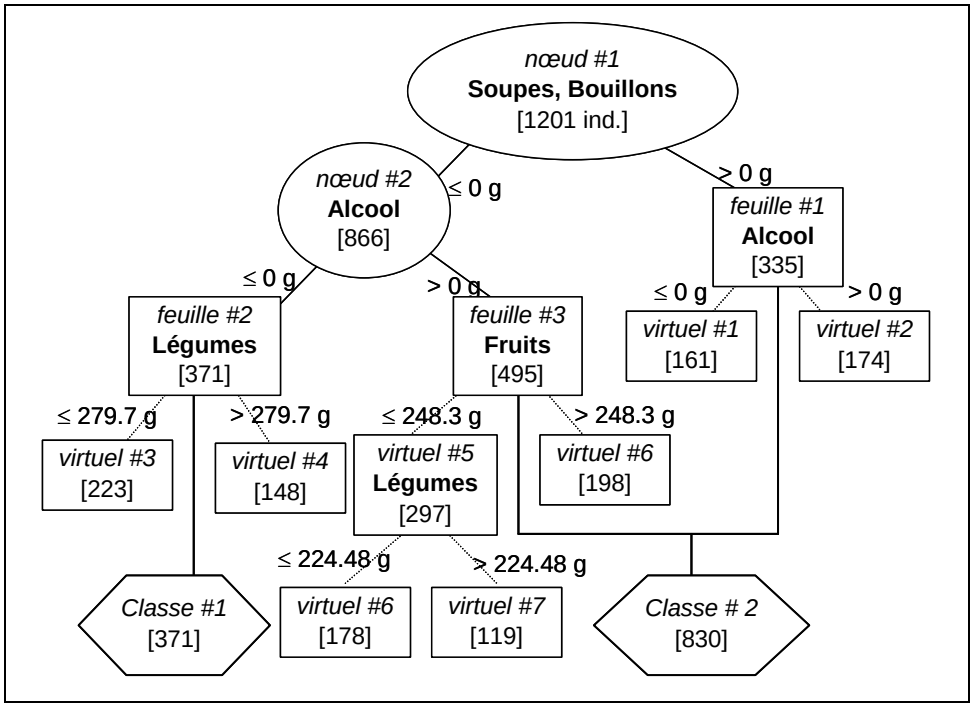


Figure 5-2 Graphe de supervision factorielle

5.2.4 Discussion

Les résultats que nous avons obtenus par cette méthode nécessitent d'autres analyses afin de les valider. Néanmoins, ils sont encourageants dans le sens que les nutritionnistes les jugent cohérents. La prochaine étape va consister à analyser la totalité des centres français pour identifier les régimes alimentaires usuels au sein d'un centre particulier ainsi qu'à l'ensemble des centres.

Nos travaux vont s'orienter vers l'implémentation du modèle général de RECPAM qui permettra de prendre en compte la présence de variables exogènes comme par exemple la variable « centre ». En outre, pour la méthode de clustering, nous cherchons à la comparer à d'autres approches notamment celle de Gower détaillée dans (GORDON, 1999).

5.3 Analyse de curriculum vitæ

En réponse aux transformations socio-économiques de ces dernières années, la mobilité des personnes s'est accrue, les changements d'emplois durant la vie active d'un salarié sont multiples. Le secteur du recrutement a pour mission de mettre en adéquation les offres d'emplois qu'il gère avec les réponses à

ces offres, souvent sous forme de curriculum vitæ. Ce secteur doit ainsi faire face à la croissance des offres d'une part, et à celle des CV d'autre part.

Nous exprimons les motivations qui conduisent à élaborer des méthodes automatiques de ciblage. Le Data Mining et plus particulièrement le Text Mining peuvent rendre de précieux services, 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 un grand opérateur du marché de l'emploi.

Dans un premier temps, nous donnons les caractéristiques du CV et montrons en quoi il constitue des objets complexes pour le traitement automatique. Puis, nous abordons la préparation des données. En effet, les CV 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. Après, nous décrivons des tentatives de modélisation. Celles-ci visent à construire des modèles de prédiction. Dans cette analyse, nous nous sommes limités à la catégorisation de CV de cadres et de CV généraux. Cette expérience devrait nous conduire à une large généralisation de cette approche. Le but recherché est d'arriver à mettre au point des automates capables d'apprendre à identifier des typologies de CV, de profils de candidat et/ou de poste, etc. sans être obligé de donner les mots clés spécifiques devant figurer dans le CV car ils peuvent être absents de certains CV pertinents. Nous souhaitons une approche qui repose sur le contenu sémantique. Enfin, nous établissons un bilan et dressons des perspectives.

5.3.1 Caractéristiques des curriculum vitæ

Le CV est un document *a priori* normalisé dans le sens où il est régi par des directives tant destinées à la forme qu'au contenu. En effet, le document doit dégager le profil et les compétences du candidat en renseignant généralement 4 parties (entête, cursus, expériences et divers). Cependant, ces directives évoluent fréquemment au cours du temps, ainsi certaines parties ont leur nom modifié, d'autres peuvent être supprimées ou au contraire rajoutées. De plus, les candidats modifient leur CV en adaptant ces directives à leur gré afin d'y faire transparaître leur personnalité. De ce fait, ces documents sont faiblement structurés et les informations qu'ils recèlent sont éparses.

En outre, le contenu du CV est fortement symbolique de part la multitude des sigles employés ou encore les divers diplômes renvoyant à des niveaux d'études équivalents. Mais le contenu sémantique y

est également très dense comme par exemple la description des activités réalisées par le candidat renseignant ainsi sur son niveau de compétence.

Face à la forte croissance de la quantité de CV à traiter, il est intéressant d'utiliser des méthodes de Data Mining. Ces méthodes sont donc confrontées à un document textuel complexe et faiblement structuré.

5.3.2 Recherche d'un espace restreint de descripteurs

Notre corpus de 686 CV est constitué de 1290 mots distincts. Il se divise en 2 classes fortement déséquilibrées : 66 CV de cadres et 620 généraux ; $\mathcal{E} = \{\text{cadre}, \text{général}\}$. L'exploitation des CV dans l'espace des mots n'est évidemment pas possible sans diminution préalable de ce dernier. 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 (cadre, général).

Comme nous l'avons vu au Chapitre 1 se rapportant aux travaux de (LUHN, 1958), une représentation vectorielle d'un texte d'un corpus se basant sur la fréquence des termes est une méthode usuelle. Cependant, dans le cadre de la modélisation des CV, il nous apparaît intéressant de travailler avec des concepts plutôt que les termes eux-mêmes. En effet, outre les inconvénients de la synonymie, le vocabulaire présent est très riche, e.g. divers noms de fonctions, de diplômes, et ayant une interprétation similaire.

Après les pré-traitements usuels (casse uniforme, regroupement du genre et du nombre d'un même mot, détection des groupes de mots, …), nous sélectionnons les termes étant ni trop fréquents, ni trop rares. Puis nous recodons les données en trois modalités (0, 1, 2), ainsi :

- la modalité 0 reflète l'absence d'un terme dans un CV ;
- la modalité l y représente un nombre modéré d'occurrences d'un terme ;
- la modalité 2 y signifie un nombre élevé d'occurrences d'un terme.

Pour la définition des concepts, nous faisons appel à la méthode des mots associés de (CALLON *et al.*, 1983). Cette méthode se base sur une méthode de classification hiérarchique (CHANDON et PINSON, 1981, Chapitre 5). Habituellement, la classification est contrainte afin de former des clusters composés d'un nombre fixé *a priori* de termes, et permettant ensuite la représentation graphique de leurs proxi-

mités structurelles (COURTIAL, 1995). Ne pouvant fixer le nombre d'éléments par cluster et la représentation structurelle ne nous étant que faiblement utile dans cette étape, nous ne contraignons pas la classification. CALLON *et al.* emploie une métrique se basant sur la notion de co-occurrences des termes[4] pour mettre en relief leurs relations. Leur utilisation se justifie lorsque l'on considère que deux documents sont voisins quand ils utilisent les mêmes termes. Cette mesure de similarité est nommée indice d'équivalence.

Soient t_1 et t_2 deux termes, alors leur indice d'équivalence dans l'ensemble du corpus est :

$$E_{t_1,t_2}^{corpus} = \frac{\left(C_{t_1,t_2}\right)^2}{C_{t_1} C_{t_2}}$$

avec :

- C_{t_1,t_2} : le nombre de co-occurrences de t_1 et t_2
- C_{t_1} : le nombre d'occurrences de t_1
- C_{t_2} : le nombre d'occurrences de t_2

Après pré-traitements, KODRATOFF *et al.* (Laboratoire de Recherche en Informatique – Université Paris-Sud) ont extrait, à partir du corpus des CV généraux et de façon non supervisée, 74 termes. Pour ce faire, ils ont défini des candidats-termes en se basant sur le type des mots utilisés dans le corpus à l'aide de l'étiqueteur de BRILL (1994). Puis, ils effectuent un élagage consistant afin de ne prendre en compte que les candidats-termes ayant un nombre d'occurrences au-dessus d'un certain seuil λ.

Pour définir les hypothèses de notre ontologie, nous avons appliqué une C.A.H. (Figure 5-3) comme décrite précédemment. Sa principale difficulté réside dans le choix du point de coupure du dendogramme. Dans notre cas, nous avons choisi d'explorer les premières agrégations puisque la mesure de similarité utilisée y regroupe les termes les plus fortement liés (CALLON *et al.*, 1983). Nous les avons coloriés en foncé et légendés avec les termes constituant ces agrégations. Nous y avons également ajouté les légendes d'agrégations plus tardives, jugées pertinentes (encadrées en pointillés sur la Figure 5-3).

[4] CALLON *et al.* parlent de mots-clefs, ce qui est équivalent dans le cas présent puisque nos termes ont été sélectionnés comme étant caractéristiques.

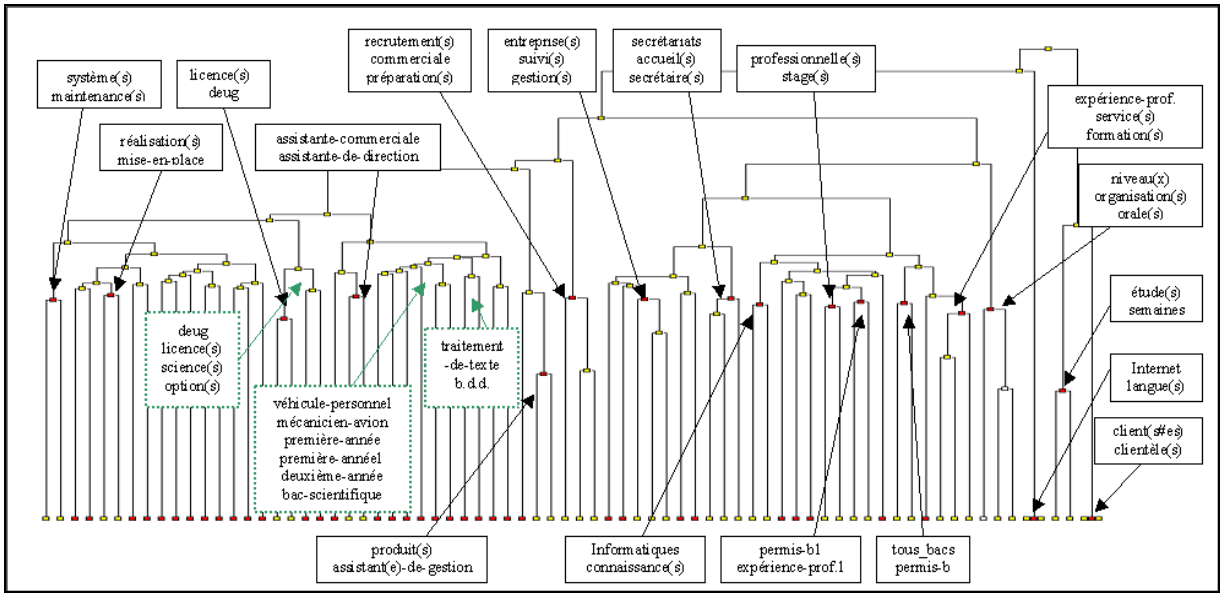


Figure 5-3. C.A.H. des termes

De ces agrégations, nous dégageons principalement deux hypothèses :

- le regroupement en fonction du niveau de compétence, e.g. :
  - deug, licence(s), science(s), option(s),
  - expérience-prof., service(s), formation(s) ;
- le regroupement selon les activités, e.g. :
  - secrétariats, accueil(s), secrétaire(s),
  - système(s), maintenance(s).

Ces hypothèses nous semblent valides dans le cas de la différenciation des CV *cadres* et *généraux* puisque par exemple, une personne munie d'un diplôme d'au moins du 2nd cycle est généralement cadre, ou encore les secteurs d'activités liés aux services sont davantage liés à la classe générale. Nous regroupons alors les différentes terminologies de diplômes selon leur cycle universitaire équivalent et selon une activité similaire (Tableau 5-1).

Après application de l'ontologie, nous sélectionnons 50 concepts suivant le $\chi^2_{\text{multivarié}}$ (Tableau 5-2), comme décrit au Chapitre 1, section 1.6.2.4. Les concepts en italique regroupent des termes liés à des zones géographiques (*Afrique, Asie, Eurasie, France*) ou à la nationalité (*étranger*) et n'ont pas été pris en compte dans la suite de l'analyse. Il en résulte donc une liste de 45 concepts.

Regroupement par diplôme		*Regroupement par activité*	
Termes	Concepts	Termes	Concepts
bep, troisieme, colleges	bepc	export, import, commande, affaires, devis	commerce
cap, seconde, premiere	lycee	econometrie, economique, marche	economie
baccalaureat, terminale	bac	statistique, classification, segmentation, discriminante, analyse, donnees, quantitati-ves, prevision, decision	statistiques
superieur, bts, iut, dut, mass, deug, deust	bac1		
iup, licence, maitrise	bac2		
dess, ecole, master, esc, hec, ingenieur, dea, docteur	bac3	…	…

Tableau 5-1 – Extraits de regroupements : par diplôme et par activité

marketing	*asie*	*france*	Service	Saisie
economie	bac3	*etranger*	Technicien	Assistance
statistiques	bac2	Gout	Standard	Centre
afrique	institut	Management	Gestion	Commercial
soft stat	stage	Stratégie	Accueil	Assurance
université	langage	Secrétaire	Sécurité	Travail
international	appliquées	Intérim	Administration	Seconde
etude	commerce	Bepc	Tenue	Classement
mathématiques	toefl	Maintenance	Préparation	Production
eurasie	sql	Agent	Qualité	Directeur

Tableau 5-2. Concepts sélectionnés par le $\chi^2_{\text{multivarié}}$

Ainsi, à l'issue de ces étapes, d'un espace de 1290 variables nous arrivons à un espace bien plus restreint car composé seulement de 45 variables. Nous avons pondéré les fréquences de ces 45 concepts en utilisant la pondération TF×IDF (1.3.8), largement utilisée car corrigeant la fréquence d'un terme, d'un concept en fonction de la proportion inverse du nombre de documents du corpus comportant le terme.

5.3.3 Prédiction des classes de CV

Le paradigme de l'apprentissage supervisé (cf. Chapitre 2) consiste à définir un modèle induit et validé à partir de données *a priori* connues. Par données *a priori* connues, nous entendons les individus (en l'occurrence les CV) dont nous connaissons leur appartenance à une catégorie d'un thème (*cadre*, ou

générale). Dans le cadre de cette étude, malgré un fort déséquilibre entre les étiquettes (10% de cadres et 90% de généraux), nous avons effectué des analyses sur la catégorisation des CV.

5.3.3.1 Méthodes

Nous avons choisi d'utiliser deux méthodes usuelles en apprentissage supervisé. 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 composé de règles d'apprentissage très informatives. Néanmoins, les méthodes à base d'arbres pouvant être confrontées au sur-apprentissage, nous comparerons nos résultats au modèle issu de l'analyse discriminante (FISHER, 1936) qui détermine le meilleur hyper-plan séparateur de nos 2 classes. En outre, les coefficients des variables nous fournissent des informations sur l'intervention des variables dans le modèle.

Pour comparer la qualité du ciblage de la classe *cadre* des 2 modèles, nous construisons une courbe de lift :

- En abscisse, le pourcentage d'individus cumulés ;
- En ordonnée, le pourcentage des scores, préalablement ordonnés de façon décroissante, cumulés. Ici, le score est la confiance liée à la classe *cadre*.

Par ailleurs, afin d'avoir une mesure plus objective de l'erreur réelle, nous utilisons le principe de la validation croisée (cf. Chapitre 2, section 2.4.3).

5.3.3.2 Modèles obtenus

Les 2 modèles construits donnent des résultats similaires. En effet, comme l'illustre le Tableau 5-3, les 10 termes les plus discriminants pour l'arbre C4.5 et l'analyse discriminante sont identiques à l'exception de 2 termes (en italique dans le tableau).

Pour illustration, la première règle de décision de l'arbre se base en première condition sur la présence du concept **statistiques** et fait ainsi apparaître une forte proportion de cadres (65% contre seulement 10% au départ). Ensuite intervient le niveau d'étude, ici **bac2** (i.e. équivalent au 2nd cycle universitaire) : son absence élimine les cadres. Enfin, le critère **softstat**, concept regroupant les différents logiciels de statistiques, permet de séparer les cadres des généraux.

En outre, les deux modèles ont également une capacité de ciblage d'un même ordre de grandeur (Figure 5-4). Ainsi, pour atteindre 90% des cadres identifiés dans le corpus, il suffit d'exploiter 10% de la population, ce qui correspond à la distribution réelle de ce corpus (9,62% de cadres). De même, les taux d'erreur en validation sont du même ordre de grandeur pour ces 2 modèles (Tableau 5-4).

	C4.5	Analyse Discriminante
1	ECONOMIE	MARKETING
2	ETUDE	ECONOMIE
3	BAC3	SOFTSTAT
4	BAC2	*STATISTIQUES*
5	*STAGE*	*INTERNATIONAL*
6	SOFTSTAT	UNIVERSITE
7	UNIVERSITE	MATHEMATIQUES
8	MATHEMATIQUES	ETUDE
9	*INSTITUT*	BAC2
10	MARKETING	BAC3

Tableau 5-3 - Les 10 termes les plus discriminants

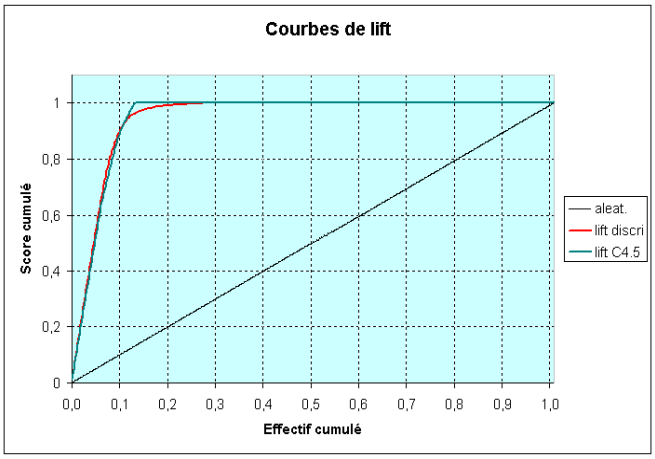


Figure 5-4. Courbes de lift

C4.5				Analyse Discriminante			
Matrice de confusion - Taux d'erreur : 9%				Matrice de confusion - Taux d'erreur : 8%			
STATUT	CADRE	GENERAL	TOTAL	STATUT	CADRE	GENERAL	TOTAL
CADRE	35	31	66	Cadre	36	30	66
GENERAL	32	588	620	Général	24	596	620
TOTAL	67	619	686	TOTAL	60	626	686

Tableau 5-4 - Matrices de confusion issues d'une 5 validation-croisée

Cependant, du fait du fort déséquilibre entre les classes, il est nécessaire d'étudier les performances partielles du classifieur (cf. Tableau 5-5). Nous remarquons dès lors les difficultés de prédiction concernant la classe *Cadre*.

	C4.5	Analyse Discriminante
ρ_{Cadre}	53%	55%
π_{Cadre}	52%	60%
$\rho_{\text{Général}}$	95%	96%
$\pi_{\text{Général}}$	95%	95%

Tableau 5-5 - Rappels et précisions des classes *cadre* et *général* selon les modèles

5.3.4 Discussion

L'intérêt majeur de ce travail est la mise en évidence d'un contenu informationnel au sein d'un corpus de CV. En effet, les deux modèles élaborés tendent vers des résultats similaires, tant du point de vue des termes intervenant dans les modèles que des caractéristiques de ciblage.

Néanmoins, la médiocrité des résultats en rappel et précision de la classe *Cadre* nous incite évidemment à ne pas exploiter tels quels ces modèles. Ces résultats devraient sensiblement s'améliorer avec un corpus comportant un nombre de CV de cadres plus conséquent. En effet, lors des 5 validations-croisées, il n'est utilisé qu'environ une dizaine de CV de cadres, ce qui semble être largement insuffisant dans ce cas.

D'un point de vue informatif, les résultats semblent intéressants et encourageants, mais nécessitent l'intervention d'un spécialiste du recrutement pour une évaluation plus objective. Cependant, d'un point de vue qualitatif, ces résultats ont besoin d'être améliorés. Pour cela, nous proposons l'utilisation d'un corpus plus conséquent et une mise en œuvre de stratégies d'apprentissage donnant un poids non symétrique aux erreurs.

Enfin, dans cette étude nous avons exploité seulement le contenu des CV (e.g. la terminologie), mais ces derniers sont constitués d'une structure très singulière (entête, cursus, expérience, divers et lettre de motivation), et il serait fort intéressant d'être en mesure d'exploiter cette information pour affiner la modélisation du corpus, et par-là même, la qualité des modèles obtenus.

5.4 Analyse de compte-rendus médicaux

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 ou 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 (ACR) adaptée pour la France par l'Agence Nationale d'Accréditation et d'Evaluation en Santé (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'information 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 cette section, 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 ces travaux ont été conjointement menés avec le Centre Léon Bérard, pôle important de la recherche contre le cancer en France.

Dans un premier temps, nous donnons les particularités des CR de mammographies et montrons en quoi ils constituent des objets complexes pour le traitement automatique. Puis, nous abordons la prépa-

ration 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. Ensuite, 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 : $\mathcal{E} = \{\text{bénin}, \text{malin}\}$. Enfin, nous établissons un bilan et dressons des perspectives pour l'amélioration et l'exploitation de nos résultats.

5.4.1 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

Tableau 5-6 Statistiques élémentaires du corpus CR

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 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 5-6).

5.4.2 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).

Contrairement au corpus des CV, nous n'avons pas cherché à mettre en place une ontologie. En effet, le vocabulaire étant très précis et les nuances des termes étant volontaires, il nous semblait inopportun de les regrouper à travers un réseau sémantique.

Par contre, les différentes flexions des termes sont (majoritairement) issues des règles grammaticales et non d'une nuance ou d'une variation du sens des termes. Ainsi, nous avons extrait les pseudo-racines à l'aide de l'outil (SNOWBALL) développé par PORTER M., comme décrit au Chapitre 1, section 1.3.4.1.

Nous avons également fait le choix de rendre homogène la casse du texte et de supprimer les mots-outils, désirant nous concentrer sur les termes médicaux à proprement parler. En outre, nous avons également écarté les opérateurs de négation. Ce choix peut paraître surprenant car ces derniers sont fortement présents au sein de ce corpus, où les documents sont fréquemment rédigés à l'aide de phrases négatives (e.g. « Il n'y a pas de zone suspecte »). Nous interprétons les phrases négatives comme la description d'un fait certain. A l'opposé, les phrases descriptives affirmatives (e.g. « Nous pouvons voir ») nous paraissent traduire la difficulté de la lecture du cliché, où le radiologue est plus réservé dans son propos. Nous pensons donc qu'ici, les opérateurs de négations ont un rôle fortement stylistique plutôt que sémantique. Ces conclusions ont été étayées par notre tentative de prendre en compte ces termes à l'aide de notre méthode de construction de variables synthétiques contextuelles. Comme

nous l'avons décrit au Chapitre 1, ces expérimentations n'ont pas donné de résultats probants puisque ces variables n'ont pas été sélectionnées à travers la méthode du $\chi^2_{\text{multivarié}}$.

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			

Tableau 5-7 Liste des 100 termes sélectionnés par le $\chi^2_{\text{multivarié}}$

Enfin, pour dernier pré-traitement, nous avons supprimé les accents des termes de notre corpus. En effet, le corpus contient beaucoup d'erreurs dans l'utilisation des accents. Par exemple pour le terme *hétérogène*, nous avons recensé 5 orthographes différentes ! Par ailleurs, l'utilisation d'un correcteur orthographique nous paraissait coûteux face aux contraintes de temps. Cependant, seule une étude comparative entre l'utilisation des accents (mots correctement orthographiés) et sans permettra de trancher la question.

A partir de cette nouvelle représentation du corpus, 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 davan-

tage, nous sélectionnons alors 50 termes par modalité (bénin, malin) selon le $\chi^2_{\text{multivarié}}$ (Tableau 5-7), que nous pondérons par le TF×IDF (1.3.8), pour les mêmes raisons qu'avec le corpus des CV.

Pour une interprétation plus aisée, nous avons regroupé *a posteriori* ces termes en huit catégories (sur le Tableau 5-7, 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 5-7, 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.

5.4.3 Modélisations des compte-rendus

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

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

Comme nous l'avons déjà vu, 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 méthode, l'analyse discriminante (FISHER, 1936), 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) (cf. Chapitre 2, section 2.2), 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 cosinus et pondéré les votes proportionnellement à l'inverse de leur distance.

Pour valider la capacité de généralisation du modèle afin de déterminer la confiance que nous pouvons accorder aux différentes informations ainsi obtenues, nous les évaluons en validation croisée. 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.

Les résultats obtenus à partir des 3 modèles utilisés à la suite d'une 10 validation croisée sont résumés dans les Tableau 5-8 et Tableau 5-9. 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 a un impact quasi équivalent.

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

Tableau 5-8 Taux d'erreur en validation croisée des 3 modèles

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

Tableau 5-9 Taux de rappel et précision des 3 modèles

5.4.4 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, tend 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 supposées 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 commune 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. Cette approche, appelée *stacking* (WOLPERT, 1992) consiste à construire un méta modèle à partir des prédictions de chacun des 3 modèles. Ce méta modèle tenterait de trouver la meilleure 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. En considérant que chaque modèle est un attribut prédictif, 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 5-10 et Tableau 5-11.

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

Tableau 5-10 Taux d'erreur en validation croisée du méta modèle

	Rappel		Précision	
Modèle	Bénin	Malin	Bénin	Malin
AD	87%	79%	84%	82%

Tableau 5-11 Taux de rappel et précision 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 rappel et précision sont supérieurs à toutes les précédentes valeurs.

Dans cette application, 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 ici 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 asymétriques 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*.

5.5 Application à la catégorisation multilingue

Dans cette application, nous proposons des solutions pour étendre la catégorisation de textes aux corpus multilingues. Bien évidemment, de nouvelles contraintes sont introduites comme la reconnaissance automatique de la langue et sa traduction automatique. Notre approche peut paraître naïve, mais elle a le mérite d'une part d'être, à notre connaissance, la première solution automatique proposée et d'autre part d'être opérationnelle à l'instar de nos premières expérimentations sur un corpus d'articles de journaux allemands, anglais et français.

Comme dans le cadre usuel de la catégorisation monolingue, la phase d'apprentissage supervisé s'effectue à l'aide d'un corpus d'apprentissage étiqueté et rédigé dans une langue donnée $\mathcal{L}_{app}$. Dans le cadre multilingue, l'inférence n'est possible pour un texte rédigé dans une langue quelconque, qu'à partir du moment où un traducteur automatique de cette langue vers $\mathcal{L}_{app}$ est disponible. Dans un esprit d'indépendance maximale face aux traitements linguistiques, coûteux et spécifiques à chaque langue, nous excluons d'office les méthodes utilisant de manière explicite des informations spécifiques à chaque langue.

Bien entendu, de même que pour la catégorisation de textes monolingues, le texte à classer doit appartenir au même domaine que les textes utilisés lors de l'apprentissage. On ne saurait, par exemple, essayer de classer un article scientifique à partir d'un modèle construit sur un ensemble d'apprentissage constitué d'articles de journaux à scandale.

Cette section est organisée de la manière suivante : dans un premier temps, nous exposons notre approche pour étendre la catégorisation de textes au cas multilingue. Puis, nous l'appliquons sur un exemple réel de catégorisation de journaux. Enfin, nous discutons des résultats obtenus, et nous tentons de mettre en perspective notre démarche afin de la faire évoluer.

5.5.1 Méthodes pour la catégorisation de textes multilingue

5.5.1.1 Nouveau cadre pour la catégorisation multilingue

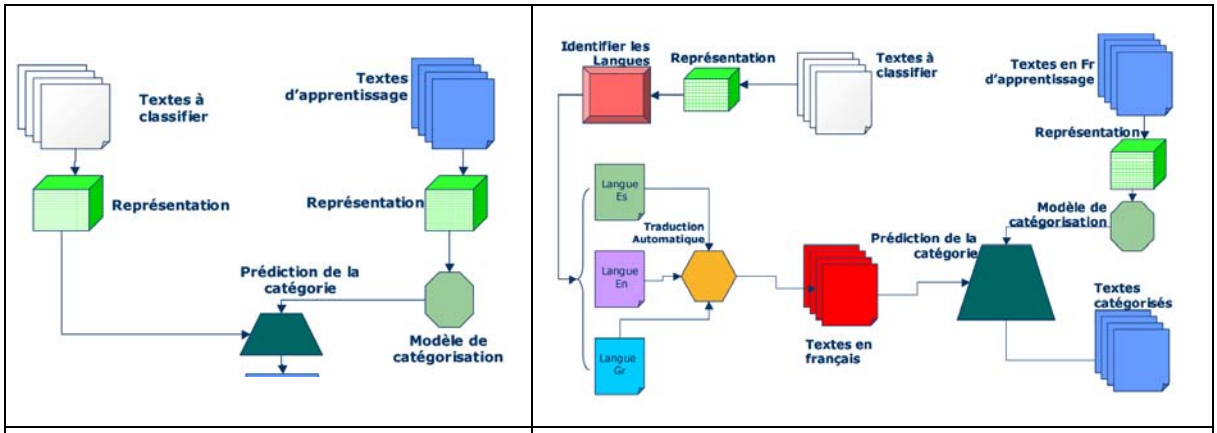
Dans le cadre de la catégorisation de textes multilingue, le processus comporte deux nouvelles exigences. Tout d'abord le corpus de textes étiquetés utilisé pour l'apprentissage doit être disponible dans

une langue $\mathcal{L}_{app}$ donnée. Ensuite, la langue de chaque nouveau texte à classer doit d'abord être déterminée avant de pouvoir lui associer son étiquette.

Pour répondre à ces nouvelles contraintes, nous aménageons le processus de catégorisation exposé au Chapitre 2 et résumé dans la Figure 5-5-a. La phase d'apprentissage n'est pas modifiée, en revanche la phase de classement comporte deux étapes supplémentaires (Figure 5-5-b) :

1. Nous devons tout d'abord détecter la langue dans laquelle le texte *i* à classer est rédigé ;
2. Si la langue est reconnue par le traducteur, le texte est traduit vers la langue $\mathcal{L}_{app}$ et il devient $\tilde{i}$.

Nous recherchons alors les occurrences des termes $\left(x_i^1,\ldots,x_i^p\right)$ dans $\tilde{i}$ afin de pouvoir appliquer le modèle prédictif φ et ainsi associer son appartenance aux différentes étiquettes de $\mathcal{E}$.



(a) Catégorisation de textes monolingue | (b) Catégorisation de textes multilingue

Figure 5-5 Schémas généraux de catégorisations mono et multilingue

5.5.1.2 Détection de la langue du texte à classer

Il est important de détecter avec précision la langue dans laquelle le texte à classer est rédigé, car une erreur à ce niveau voue à l'échec les étapes suivantes.

Il existe deux familles d'approche dans l'identification de la langue : linguistique ou statistique. En cohérence avec notre choix d'indépendance maximale face aux traitements linguistiques, nous avons privilégié l'approche statistique. Cette dernière capture automatiquement certaines régularités statistiques des langues. Comme descripteurs, nous avons utilisé les 3-grammes, séquences de 3 caractères consécutifs (cf. Chapitre 1, section 1.3.4.1), extraits du texte à classer. TEYTAUD et JALAM (2001) ont

montré qu'un texte de longueur de 100 octets permettait d'obtenir une reconnaissance d'excellente qualité, à près de 99%, et que pour des textes plus longs, la reconnaissance était parfaite.

5.5.1.3 Traduction du texte à classer

La traduction du texte à classer dans la langue du corpus d'apprentissage Lapp est également une étape primordiale. L'objectif ici n'est pas de produire un texte traduit retraçant fidèlement les propriétés sémantiques de l'original, mais de fournir un texte assurant une qualité de classement suffisante. Il est évident que le résultat obtenu dépendra du traducteur utilisé, une des perspectives immédiates de notre travail actuel consistera à analyser de manière approfondie le traducteur pour évaluer son efficacité lors de la catégorisation de textes.

Nous avons utilisé un traducteur en ligne disponible sur Internet (SYSTRAN). Ce traducteur n'est certainement pas le meilleur, mais puisqu'il est publiquement disponible, nous aurons la possibilité par la suite de comparer nos résultats avec d'autres études. Il nous semble que l'utilisation d'un meilleur traducteur devrait améliorer la catégorisation des textes.

5.5.2 Application sur le corpus CLEF

5.5.2.1 Constitution du corpus

La réalisation d'un corpus de catégorisation nécessite un investissement considérable. Le recueil des données à proprement parler est relativement aisé et grandement facilité par l'utilisation d'Internet. Cependant, la difficulté principale est d'assigner un ensemble d'étiquettes à chacun des textes recueillis. Cette étape est d'autant plus importante que sa qualité influe directement sur les performances du classifieur. Il est donc préférable de confier cette tâche à des experts des sujets traités par les documents, ce qui dans notre cas nécessite un collège d'experts manipulant les diverses langues. Il faut également que chacune de ces catégories soit suffisamment représentée dans les diverses langues.

Malgré nos recherches, nous n'avons pas trouvé de corpus multilingue de textes étiquetés en classes comparables. Nous avons alors choisi d'adapter les documents proposés par les organisateurs du concours CLEF (CLEF). Les corpus de documents utilisés dans la campagne d'évaluation CLEF proviennent de différents journaux tels que le Los Angeles Times (États-Unis), Le Monde (France), ou

des dépêches de l'agence télégraphique SDA (Suisse allemande). Les documents de ces corpus sont tous extraits de l'année 1994 et les thèmes abordés sont approximativement comparables.

Ces corpus sont destinés aux tâches de la recherche documentaire monolingue, bilingue et multilingue. Les fichiers de ces corpus sont au format SGML, par exemple, le fichier *lemonde_19940604.sgml* concerne les articles apparus dans le journal Le Monde (LM) lors de la journée du 4 juin 1994. Chaque article de ce fichier contient des balises SGML décrivant son contenu (Tableau 5-12). Il est facile d'y distinguer son numéro d'identification, son titre, et son corps.

```
<DOC>
<DOCNO>LEMONDE94-000386-19940604</DOCNO>
<DOCID>LEMONDE94-000386-19940604</DOCID>
<ACCOUNT>339369</ACCOUNT>
<GENRE>RECTIF</GENRE>
<DATE>19940604</DATE>
<LMDOC>MHB</LMDOC>
<FAB>06031011</FAB>
<SUBJECTS>DEMENTI,DESSIN</SUBJECTS>
<GO21>PUBLICATION</GO21>
<NAMES>SERGUEI</NAMES>
<PUM1>QUO</PUM1>
<REFERENCE1>2-002-06</REFERENCE1>
<SEC1>IDE</SEC1>
<PAGE>13</PAGE>
<TITLE>Serguei précise</TITLE>
<TIO1>PAS DE PANIQUE A BORD</TIO1>
<TEXT>Un lecteur s'est étonné de constater qu'une publication satirique d'extrême droite, Pas de panique à bord, citait, parmi les noms de ses collaborateurs, celui du dessinateur Sergueï. Etonnement encore plus grand _ pour ne pas dire plus _ de Sergueï, celui que nos lecteurs connaissent bien et qui ne saurait être le même que son homonyme de Pas de panique à bord, s'il existe. &gt;</TEXT>
</DOC>
```

Tableau 5-12 Exemple extrait de la collection CLEF
montrant un article du journal Le Monde publié le 4 juin 1994

La taille des corpus varie fortement entre les langues, avec des volumes plus restreints pour le français. Le nombre de mots par article reste assez similaire (environ 130), avec une moyenne un peu plus élevée pour la collection anglaise (167). Par contre, la variabilité de cette longueur demeure assez forte (écart-type d'environ 120) (SAVOY, 2002).

Comme nous l'avons vu, la catégorisation de textes nécessite d'avoir un échantillon d'apprentissage composé de textes associés à leurs étiquettes (ou classes). Une première approche peut consister à utiliser les termes-index associés à chaque article comme classes d'appartenance. En général, l'indexation est une opération qui décrit et caractérise un document, ou un fragment de document, en

repérant les thèmes présents dans ce document (AFNOR, 1993). Malheureusement, les résultats de nos expérimentations utilisant les termes-index, sur les corpus français et allemand, sont médiocres : le taux d'erreur est proche de celui du hasard.

Il est difficile d'apprendre lorsque l'on utilise les termes-index proposés en tant que classes pour diverses raisons :

- Les termes-index associés aux articles français et allemand sont trop nombreux, plus de 16200 termes (et donc classes) pour le corpus français et plus de 32000 termes pour le corpus SDA allemand. Il n'y a pas eu de règle d'indexation basée sur des vocabulaires contrôlés et beaucoup de termes sont des synonymes (e.g. *mort* et *morts* ; *France* et *fr* ; *German* et *gr*).
- Les termes-index proposés ne représentent vraiment pas les thèmes abordés dans les articles ; par exemple, on associe à la dépêche du Tableau 5-12 le mot-clef *dessin* alors qu'elle parle d'un démenti.
- Les termes-index dans les dépêches de l'agence télégraphique suisse (allemand et français) ne sont pas séparés par des signes de ponctuation ; ainsi, il est difficile d'extraire les termes composés comme *conseil de sécurité*. Le corpus de Los Angeles Times, à la différence des autres corpus, ne fournit pas du tout de termes-index, ces termes-index aidant normalement à décrire le contenu d'un article.

Dans nos expérimentations, nous choisissons donc de considérer les thèmes proposés dans la campagne CLEF 2002 comme classes à prédire. Ces dernières sont très variées ; on trouve par exemple : « U.N. sanctions against Iraq », « Conflict in Palestine » ou « Leaning Tower of Pisa ». Nous avons donc travaillé sur trois corpus de langues anglaise, française et allemande : le Los Angeles Times (LAT), Le Monde (LM) et l'agence télégraphique suisse (SDA). Les thèmes utilisés dans nos évaluations sont décrits dans le Tableau 5-13.

Classes	CLEF id	CLEF topic	#LAT	#LM	#SDA
e_1	92	U.N. sanctions against Iraq	27	24	23
e_2	95	Conflict in Palestine	96	89	66
e_3	103	Conflict of Interests in Italy	10	24	70
e_4	108	Southern Yemen Secession	18	19	63
e_5	119	Destruction of Ukrainian nuclear weapons	54	33	55
e_6	122	North American car industry	27	23	6
e_7	124	Common foreign and security policy (CFSP)	32	48	24
e_8	131	Intellectual Property Rights	40	43	43
e_9	133	German Armed Forces Out-of-area	10	21	17
e_{10}	140	Mobile phones	70	95	23

Tableau 5-13 Description des catégories de la campagne CLEF

5.5.2.2 Catégorisation des articles

Au préalable de la catégorisation, la représentation des textes est une étape critique. Nous avons choisi d'utiliser les mots et les 3-, 4- et 5-grammes. Nous avons appliqué notre algorithme de sélection de termes $\chi^2_{\text{multivarié}}$ en sélectionnant 100 mots avec pour seuls pré-traitements l'uniformisation de la casse et la suppression des mots-outils. En outre, nous avons sélectionné 200 3-, 4- et 5-grammes avec pour seul pré-traitement l'uniformisation de la casse. Nous avons choisi de sélectionner moins de mots que de n-grammes puisqu'un mot est composé de plusieurs n-grammes ; après plusieurs expériences, le choix de 100 mots et 200 n-grammes apparaît comme un bon compromis conservant la structure informationnelle du corpus.

L'étude des termes sélectionnés révèle la présence importante de noms propres, tels les noms des pays ou de leurs ressortissants, ou encore les noms de personnalités. L'étude indique également certaines difficultés de traduction. Par exemple, l'expression française « téléphone portable » est traduite en anglais par « portable phone » ce qui n'a pas le sens voulu. Les noms propres ne sont pas non plus épargnés par des difficultés de traduction ; ainsi le terme français « Koweït » est laissé tel quel au lieu d'être traduit par le terme « Kuwait ».

Par ailleurs, dans le cadre de la reconnaissance de la langue, nous avons utilisé la distance du χ^2 pour identifier les trois langues français, anglais et allemand. Ces nouveaux tests confirment les résultats précédents (TEYTAUD et JALAM, 2001) : pour des textes de taille égale ou supérieure à 100 caractères le taux de reconnaissance de la langue est de 100%.

Afin de pouvoir évaluer notre processus de catégorisation de textes dans un corpus multilingue, nous avons effectué plusieurs expériences de catégorisation. Pour chacune d'elles, nous avons évalué le taux d'erreur pour toutes les configurations de représentation des textes (100 mots, 200 3-grammes, 4-grammes et 5-grammes) et des modèles utilisés (C4.5 et les 3 plus proches voisins).

Usuellement, en catégorisation de textes, un document peut appartenir à *m* classes. Cependant, dans l'application présentée ici (voir le Tableau 5-13), chaque document n'appartient qu'à une et une seule classe. Il en résulte que les rappel et précision « micro-moyen » sont identiques et égaux au taux de succès.

Nous présentons tout d'abord les résultats en catégorisation monolingue, ils ont servi de référence pour juger de la qualité de ceux obtenus dans un contexte multilingue.

5.5.2.2.1 Catégorisation monolingue

Afin d'évaluer le niveau intrinsèque de difficulté de catégorisation des articles, nous avons mesuré l'erreur de classement pour nos 3 corpus (LAT, LM et SDA) dans leur langue d'origine.

Les résultats à l'issue d'une 10-validation croisée sont présentés dans le Tableau 5-14. Nous en dégageons 4 principaux résultats :

- Il y a un apprentissage effectif puisque le taux d'erreur est largement inférieur au taux d'erreur du classifieur par défaut ;
- Il y a un avantage important pour la méthode du 3-ppv (un écart supérieur à 10 points) par rapport à C4.5 qui souffre de la fragmentation des données ;
- Le bon apprentissage du 3-ppv laisse penser que les termes sélectionnés sont pertinents ;
- Le corpus allemand est plus facile à catégoriser que les deux autres.

	10-V.C. LAT (An)		**10-V.C. LM (Fr)**		**10-V.C. SDA (All)**	
Taux d'erreur	**C4.5**	**3-PPV**	**C4.5**	**3-PPV**	**C4.5**	**3-PPV**
100 mots	16%	8%	24%	5%	15%	3%
200 3-grammes	23%	9%	20%	9%	11%	2%
200 4-grammes	16%	7%	16%	5%	8%	2%
200 5-grammes	14%	7%	16%	5%	9%	2%

Tableau 5-14 Taux d'erreur, en validation croisée, dans les langues originelles

5.5.2.2.2 Catégorisation multilingue

Comme décrite précédemment, la catégorisation multilingue nécessite des étapes intermédiaires complémentaires, chacune pouvant générer du biais au processus même de la catégorisation. En effet, si l'étape de traduction est de qualité médiocre, l'apprentissage sera difficile et les résultats de la catégorisation seront de mauvaise qualité. Par ailleurs, comme nous l'avons dit dans la section 5.5.2.2.2, les noms propres sont très présents dans les mots sélectionnés, et de ce fait nous pouvons nous demander si de bons résultats de catégorisation ne seraient pas simplement dus à ces noms propres. Ainsi, dans l'objectif de valider notre processus de catégorisation multilingue, nous évaluons d'abord l'effet de la traduction, puis l'effet potentiel des noms propres et enfin la capacité même du modèle à être appliqué sur des textes traduits.

Les effets du traducteur. Pour mesurer l'effet du traducteur sur le contenu informationnel des documents, nous avons traduit vers l'anglais le corpus français LM et le corpus allemand SDA. Nous avons appliqué le schéma d'apprentissage monolingue (voir Figure 5-5-a) et évalué l'erreur en validation croisée (Tableau 5-15). Les résultats montrent que le contenu informationnel des corpus, du moins ce qui est nécessaire à la catégorisation, est très peu dégradé par la traduction. En effet, les différences des taux d'erreur obtenus après traduction (Tableau 5-15) comparées à ceux obtenus dans la langue d'origine (Tableau 5-14) ne sont pas significatives.

	10-V.C. LM (An)		**10-V.C. SDA (An)**	
Taux d'erreur	**C4.5**	**3-PPV**	**C4.5**	**3-PPV**
100 mots	25%	6%	12%	3%
200 3-grammes	16%	6%	9%	3%
200 4-grammes	12%	6%	9%	2%
200 5-grammes	13%	5%	12%	3%

Tableau 5-15 Taux d'erreur en validation croisée des corpus traduis en anglais
LM (An) désigne le corpus français Le Monde (LM) traduit en anglais (An) ; SDA (An) désigne le corpus allemand de l'agence télégraphique suisse (SDA) traduit en anglais (An)

	Appris et appliqué sur LM (Fr)		**Appris sur LAT (An) et appliqué sur :**			
	LM (Fr)		*LM (Fr)*		*LM (An)*	
	3-PPV	**C4.5**	**3-PPV**	**C4.5**	**3-PPV**	**C4.5**
$\rho^{\mu} = \pi^{\mu}$	95%	76%	78%	12%	89%	60%
ρ^{M}	94%	68%	78%	14%	92%	55%
π^{M}	92%	71%	80%	10%	88%	52%

(a) Représentation avec 100 mots

	Appris et appliqué sur LM (Fr)		**Appris sur LAT (An) et appliqué sur :**			
	LM (Fr)		*LM (Fr)*		*LM (An)*	
	3-PPV	**C4.5**	**3-PPV**	**C4.5**	**3-PPV**	**C4.5**
$\rho^{\mu} = \pi^{\mu}$	95%	76%	78%	12%	89%	60%
ρ^{M}	94%	68%	78%	14%	92%	55%
π^{M}	92%	71%	80%	10%	88%	52%

(b) Représentation avec 200 4-grammes

Tableau 5-16 Précision et Rappel (micro et macro-moyen) du corpus LM
représenté par 100 mots (Tableau a) et par 200 4-grammes (Tableau b). Les deux tables montrent les performances obtenues en validation croisée. On applique le modèle LM (Fr) sur des textes écrits en français et les résultats obtenus sont comparés avec ceux du modèle LAT anglais sur le corpus « LM écrit en français » et sur le corpus « LM traduit en anglais »

	Appris et appliqué sur SDA (All)		Appris sur LAT (An) et appliqué sur :			
	SDA (All)		*SDA (All)*		*SDA (An)*	
	3-PPV	**C4.5**	**3-PPV**	**C4.5**	**3-PPV**	**C4.5**
$\rho^{\mu} = \pi^{\mu}$	97%	85%	20%	17%	97%	56%
ρ^{M}	93%	77%	13%	14%	95%	50%
π^{M}	95%	70%	12%	25%	94%	62%

(a) Représentation avec 100 mots

	Appris et appliqué sur SDA (All)		Appris sur LAT (An) et appliqué sur :			
	SDA (All)		*SDA (All)*		*SDA (An)*	
	3-PPV	**C4.5**	**3-PPV**	**C4.5**	**3-PPV**	**C4.5**
$\rho^{\mu} = \pi^{\mu}$	98%	92%	21%	17%	97%	54%
ρ^{M}	94%	80%	14%	14%	97%	52%
π^{M}	94%	78%	12%	25%	96%	54%

(b) Représentation avec 200 4-grammes

Tableau 5-17 Précision et Rappel (micro et macro-moyen) du corpus SDA représenté par 100 mots (Tableau a) et par 200 4-grammes (Tableau b). Les deux tables montrent les performances obtenues en validation croisée. On applique le modèle SDA (All) sur des textes écrits en allemand et les résultats obtenus sont comparés avec ceux du modèle LAT anglais sur le corpus « SDA écrit en allemand » et sur le corpus « SDA traduit en anglais »

Nous présentons seulement les résultats concernant la représentation utilisant 100 mots et celle utilisant 200 4-grammes. Le Tableau 5-16 regroupe les résultats obtenus pour le corpus LM et le Tableau 5-17 ceux de SDA. Nous en dégageons trois résultats principaux :

- Le premier concerne la viabilité de notre approche : même si le taux d'erreur s'accroît quand on passe d'un apprentissage sur les traductions anglaises au lieu des textes originaux, la qualité de prédiction surpasse largement celle du classifieur par défaut ;
- Le second concerne la faible variabilité des résultats obtenus en fonction des représentations de texte utilisées (mots ou n-grammes): on ne peut pas dire si l'une est plus robuste que l'autre ;
- La troisième concerne le faible biais introduit par les noms propres ; les Tableau 5-16 et Tableau 5-17 montrent que l'écart entre les résultats avant et après traduction sont suffisamment significatifs : nous observons pour le modèle 3-PPV un saut de plus de 10 points pour le corpus LM (Tableau 5-16) et un saut de plus de 70 points pour le corpus SDA (Tableau 5-17) ; rappelons que nous supposions que, si les résultats du modèle estimé pour l'anglais mais appliqué à des textes en français ou allemand, n'étaient pas significativement inférieurs à ceux obtenus sur ces textes tra-

duits, alors, la contribution des noms propres lors de l'étape de catégorisation serait supérieure à la contribution des mots communs traduits.

Les effets des noms propres. Pour évaluer comment les noms propres peuvent influencer les résultats de l'étape d'apprentissage, nous avons estimé un modèle à partir du corpus LAT dans sa langue d'origine (anglais) que nous appliquons directement sur les corpus LM et SDA laissés dans leur langue d'origine (respectivement français et allemand). Une hypothèse raisonnable est de dire que si les résultats ne sont pas significativement différents de ceux obtenus après traduction, alors la contribution des noms propres lors de l'étape de catégorisation est supérieure à la contribution des mots communs traduits. Pour effectuer cette comparaison, nous avons comparé les résultats de catégorisation obtenus après traduction en anglais des corpus à classer (LM et SDA), en appliquant le modèle élaboré à partir du corpus du LAT en anglais.

Par ailleurs, le taux d'erreur de C4.5 s'explique largement par la difficulté de prédire les classes e_3, e_6 et e_9 dont les taux de rappel et de précision sont nuls (cf. Annexe B). Pour e_3 et e_9 cela est dû au seuil d'élagage (fixé à 10) correspondant au nombre de textes du corpus d'apprentissage (LAT) composant ces classes (voir Tableau 5-13). Pour e_6, C4.5 produit une seule règle, basée sur la présence du mot *auto*, provenant des expressions *auto manufacter*s ou *auto shows*, dans le corpus du LAT ; or les expressions françaises *salon de l'auto* et *industrie automobile* des articles du Monde (LM) sont respectivement traduites par *car show* et *car industries* : le terme *auto* étant traduit par *car*, le terme *auto* est absent des traductions de LM ; C4.5 ne peut donc appliquer son (unique) règle apprise.

5.5.3 Discussion

Les résultats obtenus par nos modèles sur notre corpus sont encourageants. Cette section propose de discuter les différents choix effectués pour chacune des étapes du processus afin de moduler la signification de nos résultats.

La première étape du processus consiste à définir une représentation du corpus par des termes. Nous avons choisi ici la représentation basée sur les mots et celle basée sur les n-grammes. Ce dernier choix était motivé par la capacité des n-grammes à capturer aisément les structures informationnelles basiques en s'affranchissant des problèmes de séparation des mots, de coquilles et tout autre aspect linguistique. Nos expériences n'ont pas montré une différenciation marquée entre les résultats issus d'une sélection de mots et d'une sélection de n-grammes. Nous attribuons ces similarités au contrôle de qua-

lité des corpus. En effet, ces derniers sont destinés à la presse écrite, qui est exigeante envers les fautes d'orthographe et coquilles.

La deuxième étape consiste en la sélection des termes. Les co-présences des mots sélectionnés à partir du corpus du LAT et de celui du LM traduit en anglais se situent au niveau de 50%. La moitié d'entre eux est constituée de mots communs. Nous obtenons des résultats similaires lors de la comparaison des termes sélectionnés à partir du corpus du LAT et de celui du SDA traduit en anglais. Ceci montre la similitude informationnelle de nos trois corpus. Par ailleurs, la forte quantité de noms propres n'est pas un problème en lui-même puisque nous avons vu que son apport était faible. Cependant, lorsque le traducteur ne connaît pas un terme il le laisse tel quel. De plus, les noms propres ont une faible variabilité d'écriture dans les différentes langues. Ainsi, nous pensons que les noms propres empêchent l'évaluation complète de l'effet de la traduction.

Enfin, la sélection des termes est qualitativement intéressante puisqu'elle permet une séparabilité aisée de nos 10 étiquettes. Là encore, nous nous interrogeons sur la constitution de notre corpus. Le corpus CLEF contient 96 000 articles. De ces 96 000, seuls 8 000 ont été assignés à 50 sujets (étiquettes). Pour améliorer notre corpus et confirmer nos résultats, nous envisageons la généralisation en travaillant sur l'ensemble des 96 000 textes (8 000 assignés à des classes définies et les 88 000 restants assignés à une classe « autre ») ceci rendra plus difficile la séparabilité des sujets.

La dernière étape concerne l'apprentissage ; elle a montré les bons résultats du 3-ppv sur ces corpus. Ce résultat est en opposition avec la difficulté de prédiction des k-ppv dans un espace creux et/ou avec des variables non pertinentes. Nous attribuons donc ces performances à la qualité de notre espace de représentation (bon choix des descripteurs par notre méthode du χ^2_{multi}). Enfin, le C4.5 donne des résultats convenables mais est moins performant que le 3-ppv. Comme nous l'avons vu, cela est dû aux faibles effectifs de certaines classes et au paramétrage de la méthode.

L'objet de cette application est la définition d'un processus pour la catégorisation multilingue. Nous avons introduit deux nouvelles étapes par rapport au processus monolingue : la détection de la langue du texte à catégoriser et sa traduction dans la langue du corpus d'apprentissage. Nous avons illustré notre procédé par une application sur des corpus réels de journaux écrits en trois langues (anglaise, française et allemande). Nous avons décrit chaque étape de notre processus et présenté les résultats de nos expériences. Nous concluons à l'efficacité de notre approche.

Nous envisageons de perfectionner notre cadre pour la catégorisation de textes multilingue en proposant un schéma plus général dans lequel nous fusionnerons des corpus d'apprentissage traduits en une langue commune d'apprentissage. Nous espérons ainsi que les particularités propres à chaque langue ne seront plus retenues par les modèles estimés.

5.6 Ré-étiquetage dans un contexte de catégorisation de textes

Les trois applications précédentes (5.3, 5.4 et 5.5) décrivent des problèmes de catégorisation de documents textuels, l'objectif étant d'être capable d'assigner automatiquement les étiquettes à un ensemble de nouveaux textes. Usuellement, l'étiquetage de l'ensemble d'apprentissage est réalisé par un expert. Ainsi, pour un texte donné, le pré-étiquetage peut varier dépendamment de qui l'a étiqueté et quand. Dans certains cas, ce pré-étiquetage peut être considéré comme inconsistant.

Plus formellement, le processus d'apprentissage génère un modèle φ capable de produire la probabilité qu'un nouveau document a d'appartenir à chacune des étiquettes définies dans l'ensemble d'apprentissage. Cependant, si cet ensemble est inconsistant, le modèle peut être de mauvaise qualité. Pour diminuer ces effets, une possibilité consisterait à regrouper un comité d'experts convenant ensemble d'un unique étiquetage pour chacun des documents de l'ensemble d'apprentissage. En effet, le gain de cette opération serait la réduction certaine de l'hétérogénéité entre les experts. Cependant, et spécifiquement pour les grands corpus, une quantité d'efforts et de temps est considérable pour obtenir un étiquetage consensuel pour l'ensemble des textes.

Les textes pour lesquels les étiquettes associées seraient susceptibles de varier en fonction de l'expert ou du moment de l'étiquetage, peuvent être perçus comme des cas mal étiquetés ou comme du bruit. En outre, BRODLEY et FRIEDL (1996 ; 1999) ont montré que lorsque le bruit concerne seulement l'ensemble d'apprentissage alors la suppression des individus mal-étiquetés améliore considérablement les résultats de l'étape de généralisation, à condition que le ratio de bruit ne dépasse pas 20% voire 40% dans certains cas. Ce point de vue argumente que la subjectivité de l'étiquetage pénalise l'étape de généralisation.

Par ailleurs, pour le traitement de données qualitatives mal étiquetées, (WILSON, 1972 ; JOHN, 1995 ; BRODLEY et FRIEDL, 1996 ; BRODLEY et FRIEDL, 1999) ont exploré la détection et suppression de telles données tandis que LALLICH *et al.* (2002) et MUHLENBACH (2002) ont choisi la détection pour le ré-étiquetage.

Nous proposons ici une approche basée sur les graphes de voisinage, consistant à ré-étiqueter automatiquement l'ensemble d'apprentissage avant de réaliser le processus d'apprentissage. Cette étape peut être considérée comme étant un pré-traitement de nos données, l'objectif étant de réduire le bruit sur l'ensemble d'apprentissage en atténuant les conséquences d'un étiquetage subjectif. Pour ce faire, nous utilisons un processus itératif de relaxation autorisant la modification du pré-étiquetage sous réserve d'améliorer l'homogénéité de ce dernier. En d'autres termes, ce processus examine toutes les étiquettes de chacun des textes de l'ensemble d'apprentissage et change certaines d'entre-elles à condition d'obtenir une meilleur consistance (ce qui équivaut à diminuer l'inconsistance). Cette inconsistance est évaluée par un critère mathématique combinant deux notions. La première est la Similarité (S) évaluée à partir de la similarité entre l'étiquetage initial et l'étiquetage courant. La seconde est la Cohérence Locale Moyenne (CLM), calculée en fonction des étiquettes du voisinage de chacun des textes. La propriété principale recherchée peut se résumer en ces mots : « Le ré-étiquetage doit d'une part rester le plus similaire possible de l'étiquetage initial et d'autre part les étiquettes de chaque texte doivent être similaires à celles des textes de leur voisinage ».

Dans cette section, nous allons d'abord présenter les notations complémentaires nécessaires en 5.6.1. Nous aborderons ensuite en 5.6.2 la description de la technique de relaxation et ses deux critères ; puis en 5.6.3 nous la mettrons en œuvre sur la très connue collection de Reuters-21578 (ApteMod). Enfin en 5.6.4, nous discuterons des résultats ainsi obtenus.

5.6.1 Définitions et notations

Nous rappelons que l'ensemble d'apprentissage $\Omega_{App} \subset \Omega$ est composé de n_{App} documents pré-étiquettés. Nous fixant dans le cadre général où un texte peut appartenir à plusieurs étiquettes, alors $\mathrm{E} = \left(Y^1, \ldots, Y^m\right)$. Pour tout texte $i \in \Omega_{App}$ et toute variable $Y^k \in \mathrm{E}$, $y_i^k \in [0,1]$ est la probabilité d'appartenance de i à l'étiquette $e_k \in \mathcal{E}$.

Pour chaque texte i, nous associons un vecteur $P_i = \left(y_i^1, \cdots, y_i^m\right) \in [0,1]^m$ donnant la probabilité d'appartenance de i pour chacune des m étiquettes de $\mathcal{E}$. L'étiquetage initial du texte i est noté $P_i^0 = \left(y_i^1(0), \ldots, y_i^m(0)\right)$ et P_i est appelé étiquetage courant du texte i. De même, nous notons par $\mathcal{P}^0$ l'étiquetage initial de Ω_{App} et par $\mathcal{P}$ son étiquetage courant.

Nous faisons remarquer que la somme des probabilités d'appartenance pour un texte i n'est pas nécessairement égale à 1 puisque nous considérons qu'un document peut appartenir à plusieurs étiquettes. Par ailleurs, pour l'étiquetage initial la probabilité $y_i^k(0)$ qu'un texte i appartienne à l'étiquette k est égale à 1 si l'expert juge que le document appartient effectivement à cette étiquette, 0 sinon.

5.6.2 Méthode de relaxation multi-étiquettes

Les techniques de relaxation proviennent de méthodes de traitements d'images où elles ont été utilisées à l'origine pour restaurer les niveaux de gris d'images altérées. Ces techniques de restauration s'effectuent de manière itérative et s'appuient sur la valeur des pixels voisins. Nous évaluons ici si ce genre de technique peut apporter des bénéfices pour le traitement particulier de la catégorisation de documents textuels où chaque document appartient à un ensemble de catégories.

Nous définissons dans un premier temps les deux critères utilisés : la Similarité et la Cohérence Locale Moyenne. Nous détaillerons ensuite la mise en œuvre de la relaxation afin de rechercher l'optimum global.

5.6.2.1 Similarité

Le résultat de la relaxation est de modifier les probabilités d'appartenance à certaines étiquettes. En accord avec nos notations, $\mathcal{P}^0$ désigne l'étiquetage initial et $\mathcal{P}$ l'étiquetage courant. Le rôle de la propriété de similarité est de s'assurer que $\mathcal{P}$ ne s'éloigne pas trop de l'étiquetage initial. En effet, même si nos données peuvent être considérées comme mal étiquetées, il nous faut préserver « au mieux » l'information globale du corpus. Nous rappelons que notre objectif est de rendre plus cohérent l'étiquetage et non de l'éloigner de l'information qu'il peut contenir.

Pour mesurer la similarité par rapport à l'étiquetage initial, nous nous basons sur les écarts des probabilités d'appartenance pour chaque texte du corpus entre l'étiquetage initial et l'étiquetage courant :

$$S(\mathcal{P}^0, \mathcal{P}) = \sum_{i \in \Omega_{App}} \sum_{k=1}^{m} \left(y_i^k(0) - y_i^k \right)^2 \tag{5.6.1}$$

5.6.2.2 Cohérence locale moyenne

Nous avons vu au Chapitre 2 que les graphes de voisinage modélisent la similitude entre des documents représentés comme des points dans l'espace des descripteurs. En outre, l'équation (2.3.4) ré-

sume les relations d'inclusion entre les différents graphes de voisinage. L'Arbre de Recouvrement Minimal et le Graphe des Polyèdres de Delaunay nous sont faiblement informatifs en raison respectivement d'une trop faible quantité d'arêtes ou d'un trop grand nombre. En pratique, le Graphe de Gabriel et le Graphe des Voisins Relatifs (GVR) donnent des performances similaires. Ici, nous utilisons le Graphe des Voisins Relatifs.

Dans un contexte de relaxation, nous supposons que les étiquettes d'un document sont « relativement » proches des étiquettes des documents voisins. Afin de mesurer cette dissimilarité relative, nous définissons la cohérence locale entre un document et ses voisins telle que plus leurs étiquettes sont similaires, plus leur cohérence locale tend vers 0 et plus leurs étiquettes sont différentes plus la cohérence locale tend vers m. Ainsi, en généralisant à l'ensemble du corpus, nous définissons la Cohérence Locale Moyenne (CLM) comme étant la moyenne des cohérences locales de chacun des textes :

$$CLM(\mathcal{P}) = \sum_{i \in \Omega_{App}} \left[\frac{1}{|\mathcal{V}_i|} \sum_{j \in \mathcal{V}_i} \left(\sum_{k=1}^{m} \left(y_i^k - y_j^k \right)^2 \right) \right] \tag{5.6.2}$$

5.6.2.3 Définition de l'optimum global

En s'appuyant sur ces deux critères, le processus de relaxation modifie les probabilités d'appartenance aux étiquettes de telle façon à maximiser la cohérence globale moyenne tout en restant le plus proche possible de l'étiquetage initial. Mathématiquement, nous définissons dans l'équation (5.6.3) la fonction F comme la somme de la fonction de similarité S définie par l'équation (5.6.1) et de la fonction de cohérence locale moyenne CLM définie par l'équation (5.6.2). Le processus de relaxation vise à minimiser cette équation dans le but de trouver l'étiquetage optimal.

Le paramètre λ permet d'assigner la confiance envers l'étiquetage initial :

$$F(\mathcal{P}) = \lambda S(\mathcal{P}, {}^{0}\mathcal{P}) + CLM(\mathcal{P}) \; ; \; \lambda \in \mathbb{R}^{+*} \tag{5.6.3}$$

Pour démontrer que $F(\mathcal{P})$ possède un minimum unique, LARGERON (1991) se base sur la propriété que les fonctions strictement convexes admettent un minimum unique. Elle a prouvé la convexité stricte de F en démontrant que F est la composée de deux fonctions strictement convexes (S et CLM). Ainsi, F admet un minimum unique et nous notons cet étiquetage optimal $\mathcal{P}^*$. Par ailleurs, cet optimum peut être estimé itérativement par l'équation (5.6.4) (Zighed *et al.*, 1990) : à partir de

l'étiquetage obtenu à l'itération *iter*, les probabilités d'appartenance sont modifiées en fonction des valeurs des documents voisins et convergent vers l'optimum global.

$$\forall i \in \Omega_{App} \; ; y_i^k(iter+1) = \frac{\lambda \cdot y_i^k(0) + \sum_{j \in \mathcal{V}_i} \left[\left(\frac{1}{|\mathcal{V}_i|} + \frac{1}{|\mathcal{V}_j|} \right) \cdot y_j^k(iter_0) \right]}{1 + \lambda + \sum_{j \in \mathcal{V}_i} \frac{1}{|\mathcal{V}_j|}} \qquad (5.6.4)$$

$$\text{avec } k \in [1, \ldots, m] \text{ et } iter_0 = \begin{cases} iter & \text{si } j > i \\ iter + 1 & \text{si } j < i \end{cases}$$

Algorithme 5-1 Ré-étiquetage par relaxation

```
Relaxation(entrées: epsilon, lambda; entrée-sortie: y[])
Début
  Répéter
     eta ← 0;
     Pour chaque i appartenant à Omega Faire
        Pour chaque k appartenant à K Faire
          numerateur ← 0; denominateur ← 0; tmp ← 0;
          Pour chaque j appartenant au voisinage de i
             numerateur ← numerateur + P[j,k] * (1/card(voisinage de i) +
                          1/card(voisinage de j));
             denominateur ← denominateur + 1/card(voisinage de j);
          FinPour;
          tmp ← y[i,k];
          y[i,k] ← (lambda * y[i,k]+numerateur)/(1+ lambda + denominateur);
          tmp ← |tmp – y[i,k]|; eta ← eta + tmp;
        FinPour;
     FinPour;
  Jusqu'à (eta < epsilon) //Critère de convergence
Fin.
```

L'Algorithme 5-1 implémente l'équation (5.6.4). Il prend en entrée deux paramètres ε et λ. Le premier permet d'indiquer le seuil à partir duquel nous considérons qu'il y a convergence et le second détermine la confiance de l'étiquetage initial. Le paramètre d'entrée-sortie correspond à l'étiquetage du corpus.

5.6.3 Application sur la collection Reuters-21578 ApteMod

Afin d'expérimenter notre technique de relaxation dans le cadre de la catégorisation de textes et afin de rendre comparable nos résultats, nous avons choisi la collection Reuters (APTE *et al.*, 1994 ; LEWIS et RINGUETTE, 1994 ; COHEN et SINGER, 1996 ; YANG et PEDERSEN, 1997 ; YANG et LIU, 1999). Ainsi, nous décrivons succinctement la collection Reuters-21578 ApteMod, et décrivons les différents

paramétrages effectués pour la réalisation de nos expérimentations. Ensuite, nous présentons les résultats obtenus.

5.6.3.1 La collection Reuters-21578 ApteMod

Comme expliquée dans (YANG et LIU, 1999), cette collection est une version récente de ce corpus obtenu d'une part en supprimant des document non étiquetés comme dans Reuters-22173 (YANG et PEDERSEN, 1997), et d'autre part en sélectionnant les catégories ayant au moins un document dans l'ensemble d'apprentissage et un dans l'ensemble de test. Il en résulte $m = 90$ catégories avec $n_{App} = 7769$ documents pour l'ensemble d'apprentissage et $n_{Test} = 3019$ pour l'ensemble de test[5]. Le vocabulaire utilisé est riche : environ 17 000 mots après désuffixation et suppression des mots outils. Le nombre de catégories par document est hautement déséquilibré : 1,3 en moyenne avec un maximum de 14 catégories pour quelques documents. Pour plus de détails sur cette collection, voir (YANG et LIU, 1999).

5.6.3.2 Paramétrage

Afin de préparer le corpus pour la catégorisation, nous appliquons les mêmes pré-traitements que dans (YANG et PEDERSEN, 1997) : les documents sont convertis en minuscule, les mots outils sont supprimés et nous avons appliqué l'algorithme de désuffixation de PORTER (1980).

Nous appliquons la méthode de sélection d'attribut du $\chi^2_{\max}$ (cf. section 1.6.2.3 du Chapitre 1) et sélectionnons ainsi les 1000 meilleurs termes ; en accord avec nos notations, nous avons alors $p = 1000$. Le choix de 1000 est un compromis entre l'obtention des meilleurs résultats (obtenus pour 2000 variables par (YANG et PEDERSEN, 1997)) et l'économie en temps de calcul. La pondération des occurrences se base sur le TF×IDF (1.3.8).

Après construction du graphe de voisinage sur l'ensemble d'apprentissage, nous exécutons le processus de relaxation. Nous avons utilisé la métrique cosinus (RAJMAN et LEBART, 1998) pour l'élaboration du graphe des voisins relatifs. Enfin, nous avons fixé différentes valeurs de λ en fonction de la confiance que nous associons à la qualité du pré-étiquetage.

[5] Le découpage de cette collection est disponible sur http://www-2.cs.cmu.edu/~yiming/

Pour pouvoir mesurer l'impact de la relaxation, nous avons comparé l'évolution de l'efficacité d'un classifieur sur le même ensemble test, à partir des données brutes puis à partir des données relaxées. Comme classifieur, nous avons utilisé les k Plus Proches Voisins (kPPV) (MITCHELL, 1997) puisque c'est l'un des meilleurs algorithmes pour la classification de textes et en particulier sur cette collection comme montré dans (YANG et LIU, 1999). Nous avons fixé *k* à 10.

5.6.3.3 Résultats expérimentaux

Pour chaque document le kPPV retourne un vecteur contenant un score d'appartenance pour chaque catégorie. Voulant davantage mesurer l'impacte de la relaxation que l'efficacité du classifieur à proprement parler, nous utilisons la méthode de la précision moyenne des 11 points décrite au 2.4.5.2 et utilisée par (YANG et PEDERSEN, 1997) sur ce même corpus : 11 seuils de rappels sont définis de 0% à 100% par pas de 10% puis la valeur de la précision pour chacune de ces valeurs est calculée. La moyenne de ces 11 valeurs de précision estime la capacité de catégoriser un document. La moyenne de ces résultats obtenus pour les différents textes de l'ensemble de test permet d'évaluer la capacité globale du classifieur sur ce corpus.

Dans le cadre de nos expérimentations, nous avons choisi la collection Reuters-21578 ApteMod puisque cette dernière a fait l'objet d'une attention particulière quant à sa qualité. Ainsi son étiquetage initial peut être considéré comme idéal, et nous considérons le score de précision moyenne obtenu à partir de ce corpus (n'ayant subi aucune déformation de notre part), comme étant notre score de référence. Même si cet étiquetage nous semble idéal dans le sens où il a été contrôlé par plusieurs experts, nous appliquons notre étape de relaxation dans le but d'évaluer la stabilité des résultats avant et après relaxation. Logiquement, le processus de relaxation ne devrait modifier qu'un faible nombre d'étiquettes puisque nous allons associer à λ une valeur suffisamment élevée traduisant notre confiance envers la qualité de l'étiquetage initial. D'un autre côté, afin de nous retrouver dans une situation où l'étiquetage contient du bruit, nous allons modifier aléatoirement la valeur de 10% des étiquettes de l'ensemble d'apprentissage. Dans ce cas, nous attendons que les résultats obtenus après relaxation se rapprochent sensiblement du score de référence.

En conséquence, nous avons effectué 4 expériences soumises aux mêmes conditions d'apprentissage et de tests, mais en utilisant 4 pré-traitements différents pour l'ensemble d'apprentissage :

1. **Aucun pré-traitement :** l'ensemble d'apprentissage est inchangé ;
2. **Relaxation :** nous appliquons l'étape de relaxation sur l'ensemble d'apprentissage en fixant λ à 10 en raison de notre forte confiance envers l'étiquetage initial ;
3. **Bruitage et Relaxation :** nous utilisons l'ensemble d'apprentissage bruité à hauteur de 10% auquel nous appliquons une étape de relaxation en fixant lambda à 0,25 en raison de notre faible confiance envers le pré-étiquetage bruité ;
4. **Bruitage :** nous utilisons l'ensemble d'apprentissage bruité.

Evidemment, nous utilisons le même corpus bruité pour les étapes 3 et 4. Le bruitage a été effectué en inversant la valeur d'une étiquette sélectionnée aléatoirement, i.e. si sa valeur était 0, elle a été remplacée par 1 et réciproquement. Nous avons sélectionné aléatoirement 10% des valeurs d'un total de 7 769 documents fois 90 catégories, soit un changement de 69 921 valeurs. En d'autres termes, nous avons changé en moyenne 9 étiquettes par textes.

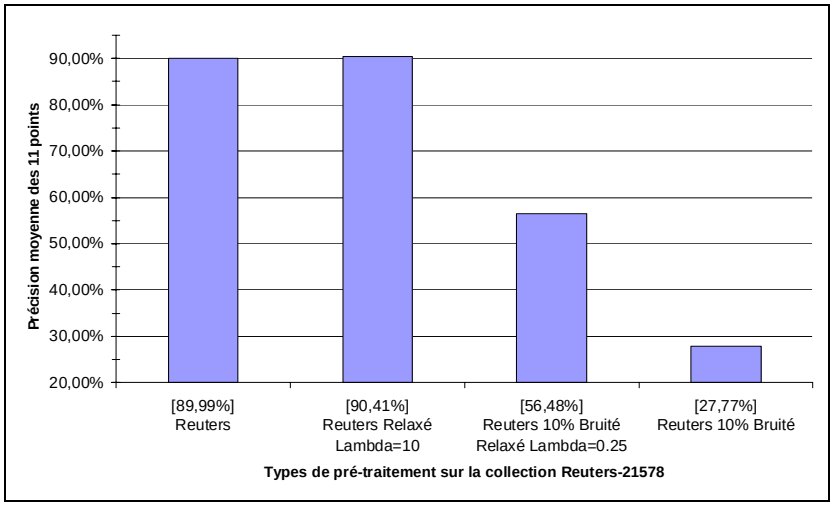


Figure 5-6 Précision moyenne du kPPV en fonction des pré-traitements

La Figure 5-6 montre la précision moyenne obtenue pour chacune des expériences effectuées. Comme illustrée, la première expérience (sans pré-traitement) retourne un score proche des 90%, score se trouvant en adéquation avec d'autres expériences effectuées sous des conditions similaires comme dans (YANG et LIU, 1999). La deuxième expérience retourne un score similaire à la première, i.e. 90%. La troisième expérience renvoie un score proche de 60%, tandis que la dernière a un score de l'ordre de 30%, soit, comme attendu, le plus mauvais score.

5.6.4 Discussion

Ces expérimentations montrent que la relaxation peut apporter des gains substantiels. En effet, son utilisation sur le corpus bruité apporte un gain spectaculaire de l'ordre de 30%. De plus, les résultats sont similaires avant et après relaxation de la collection originale, ce qui montre l'efficacité de la fonction de similarité définie par l'équation (5.6.1).

Nous en déduisons que ce type de préparation des données apporte des gains sur des collections bruitées ou non. Cependant, le type de bruit utilisé ici n'est pas réellement comparable avec la subjectivité de l'étiquetage par un expert puisque ce dernier est en principe capable d'argumenter ses choix. Ainsi, nous travaillons à tester notre méthode sur des corpus dont l'étiquetage est réputé douteux. Cependant, cela risque de provoquer des problèmes d'évaluation : en supposant que nous travaillons sur un tel corpus, comment s'assurer de la validité de l'étiquetage de l'ensemble de test, si ce n'est qu'en demandant à des experts d'évaluer les réponses ? Pour pallier cet inconvénient, nous cherchons également à mieux modéliser les divergences d'étiquetage entre des experts. Une hypothèse raisonnable semble être que pour un texte jugé par deux experts, les étiquettes qui diffèrent sont sémantiquement proches. Des groupes d'étiquettes peuvent alors être définis formant ainsi une hiérarchie d'étiquettes. Dès lors, nous pouvons inclure du bruit sur l'étiquetage initial en modifiant une étiquette par le choix aléatoire d'une des autres étiquettes appartenant au même groupe.

L'étape de relaxation peut paraître coûteuse puisque l'étape élaborant le graphe de voisinage de l'ensemble d'apprentissage a une complexité en $\mathrm{O}\left(n^3\right)$. Néanmoins, cette opération ne doit être effectuée qu'une seule fois puisqu'elle ne s'applique que sur l'ensemble d'apprentissage. Et cela est certainement moins coûteux que d'exiger un étiquetage manuel de grande qualité. Enfin, nous avons effectué ces mêmes expériences en utilisant divers types de graphes de voisinage mais au final les différences étant infimes nous ne les présentons pas.

Nous avons utilisé dans cette application une méthode de relaxation permettant de rendre plus consistant l'étiquetage initial de l'ensemble d'apprentissage d'un corpus de textes en modifiant automatiquement cet étiquetage. Nous avons en outre montré que la relaxation alliée à l'analyse de contenu des textes apporte des bénéfices évidents dans le cadre de la catégorisation de documents. Ainsi, cette méthode permet d'utiliser des corpus ayant un étiquetage douteux. De plus, cette technique peut être appliquée dans des domaines où l'étiquetage est hautement subjectif. Cependant, nous la considérons non applicable dans les domaines où une observation est objective. Par exemple, considérons le do-

maine de recherche d'analyse de survie, appliquer ce type d'outils reviendrait à considérer qu'un cas constaté mort peut être considéré vivant et inversement.

Nos futurs travaux vont consister à appliquer cette méthode sur des corpus réellement bruités et de mettre en œuvre des méthodes d'évaluation des gains ainsi apportés. Enfin, il nous semble également intéressant d'étudier comment utiliser ce type d'outils afin d'optimiser directement les fonctions de coûts des classifieurs.

5.7 Conclusion

Dans ce chapitre nous avons traité des applications concrètes liées à l'analyse et l'exploitation de données complexes. Nous avons vu que cela pouvait nécessiter l'extension ou du moins l'adaptation des méthodes employées, par exemple l'extension de RECPAM au cadre multivarié et l'adaptation de la catégorisation au cadre multilingue. Nous décrivons également notre démarche face à l'analyse de deux corpus de documents textuels atypiques : l'analyse de curriculum vitæ puis l'analyse de compte-rendu médicaux. Enfin, à travers la méthode de re-étiquetage que nous avons mise en œuvre, nous avons directement exploité la notion de contextualisation que nous avons à plusieurs reprises décrit au sein de cette thèse.

Toutes ces analyses n'ont pas apporté la même satisfaction en termes de performances. Les perspectives liées à leurs améliorations sont nombreuses et largement discutées dans ce chapitre.