PoBOC : un algorithme de "soft-clustering". Applications à l'apprentissage de règles et au traitement de données textuelles.

Guillaume Cleuziou, Lionel Martin, Christel Vrain
LIFO, Laboratoire d'Informatique Fondamentale d'Orléans
Rue Léonard de Vinci B.P. 6759
45067 Orléans cedex 2 - FRANCE
{cleuziou,martin,cv}@lifo.univ-orleans.fr
http ://www.univ-orleans.fr/SCIENCES/LIFO/

Résumé. Nous décrivons l'algorithme PoBOC (Pole-Based Overlapping Clustering) qui génère un ensemble de clusters non-disjoints (ou "soft-clusters") présentés sous forme d'une hiérarchie de concepts à partir de la seule matrice de similarités sur les données considérées. Nous évaluons l'approche sur deux situations d'apprentissage : la classification par apprentissage de règles et l'organisation de données plus complexes et peu structurées telles que les données textuelles.
La validation des méthodes de clustering est une étape difficile résolue le plus souvent par une évaluation d'experts. Les deux applications proposées permettent de valider la méthode d'organisation selon deux points de vue : d'une part quantitativement en évaluant l'influence de la méthode pour la classification, d'autre part en permettant une analyse "humaine" du résultat dans le cas des données textuelles. Nous mettons en évidence l'intérêt de PoBOC comparativement à d'autres approches d'apprentissage non-supervisé.

1 Introduction

Le clustering consiste à organiser les données de manière à regrouper les objets les plus similaires et à séparer ceux qui se ressemblent le moins. De nombreux algorithmes ont été proposés dans divers domaines d'application : la reconnaissance de formes [Jain *et al.*, 2000], la classification par apprentissage non-supervisé [Agrawal *et al.*, 1992], la segmentation d'images [Pham et Prince, 1998] ou encore le regroupement de données textuelles [Baker et McCallum, 1998]. Dans cette étude nous distinguons les algorithmes de regroupement "dur" (ou *hard-clustering*) qui renvoient un ensemble de groupes disjoints, des méthodes de regroupement "flou" (ou *fuzzy-clustering*) pour lesquelles sont renvoyés un ensemble de foyers (points dans l'espace ou objets réels) ainsi qu'une matrice d'appartenance floue. Pour ces deux types d'approches, on parle d'algorithmes de regroupement "objet" lorsque les données sont décrites par des attributs, par opposition au regroupement "relationnel" basé sur une matrice de similarités. De nombreux algorithmes ont été proposés dont certainement le plus connu est c-means (algorithme de partitionnement objet dur) et ses variantes : c-medoids (variante relationnelle), fuzzy-c-means (variante floue) et fuzzy-c-medoids (variante relationnelle floue) [MacQueen, 1967]. Ces algorithmes, avec les méthodes agglomératives hiérarchiques

SAHN [Sneath et Sokal, 1973], est couramment utilisée comme base de comparaisons afin de démontrer la performance d'une nouvelle approche.

Dans cette étude, nous proposons l'algorithme PoBOC (Pole-Based Overlapping Clustering) appartenant à une classe appelée “soft-clustering”. PoBOC présente alors la particularité d'allier les avantages des approches dures et floues à savoir la simplicité de représentation des données et la souplesse d'organisation. En effet, le “soft-clustering” est en quelque sorte un compromis entre le “hard-clustering” auquel on reproche souvent qu'un objet ne puisse appartenir qu'à une seule classe, et le “fuzzy-clustering” qui demande un post-traitement afin de pouvoir exploiter une réelle organisation en classes. Il existe très peu d'algorithmes de “soft-clustering” actuellement. On note par exemple, les méthodes de regroupement par construction de pseudo-hiérarchies (ou pyramides) [Diday, 1986]. Cette technique permet une visualisation assez riche de l'organisation de l'ensemble d'objets, cependant par définition même d'une pseudo-hiérarchie, à un niveau donné de l'arbre, un cluster s'intersecte avec au plus deux autres groupes. D'autres approches de “soft-clustering” envisagées traitent de domaines d'applications précis : par exemple l'algorithme WBSC [Lin et Kondadadi, 2001] est lié aux données textuelles et plus particulièrement au regroupement de documents.

L'algorithme PoBOC est évalué sur deux applications distinctes : dans un premier temps il est utilisé pour séparer les instances d'une classe en sous-groupes généralisables par une règle ; on montre alors que les règles ainsi obtenues sont de meilleure qualité que celles induites par d'autres algorithmes de regroupements traditionnels. La deuxième application concerne le regroupement de mots issus du langage naturel : à nouveau les recouvrements entre classes, autorisés par PoBOC, permettent d'aboutir à une organisation des mots d'avantage représentative de leur utilisation que pour d'autres méthodes d'organisation.

Plus généralement, l'objectif de ce travail est double : montrer que l'organisation des données peut permettre d'améliorer les performances d'un processus d'apprentissage et que l'organisation par une méthode de “soft-clustering”, en particulier par l'algorithme PoBOC, est préférable sinon déterminante.

L'article est organisé comme suit : la section 2 présente l'algorithme PoBOC, l'organisation des données pour l'apprentissage de règles de classification est présentée en section 3 et le regroupement de termes, en section 4. Enfin l'étude est complétée par une analyse de PoBOC.

2 PoBOC : regroupement basé sur les pôles

2.1 L'algorithme de regroupement PoBOC

L'algorithme PoBOC (Pole-Based Overlapping Clustering) prend en entrée une matrice de similarités et construit une hiérarchie de concepts dans laquelle chaque objet peut appartenir à un ou plusieurs concepts. Il se décompose en quatre étapes : (1) la recherche et la définition de *Pôles*, (2) la construction d'une matrice d'appartenance de chaque objet à chacun des pôles, (3) l'affectation des objets à un ou plusieurs pôles et enfin (4) l'organisation des groupes obtenus en une hiérarchie.

La notion de *pôle* est centrale dans l'algorithme PoBOC, il s'agit de rechercher

dans l'ensemble des objets $X = \{x_1, \ldots, x_n\}$ des zones homogènes formées de plusieurs objets (sous-ensembles de X), et situées plutôt en périphérie. La recherche des pôles s'effectue sur le graphe des similarités :

Définition 2.1 *Soient $X = \{x_1, \ldots, x_n\}$ et S une matrice de similarités définie sur $X \times X$ à valeurs dans $[-1, 1]$. On appelle* **graphe de similarités** *et on note $G_S(X, V)$ le graphe ayant X pour ensemble de sommets et V pour ensemble d'arêtes tel que :*

$$(x_i, x_j) \in V \text{ ssi } s(x_i, x_j) \geq max\{\frac{1}{n} \sum_{x_k \in X} s(x_i, x_k), \frac{1}{n} \sum_{x_k \in X} s(x_j, x_k)\} \qquad (1)$$

Les deux termes dans *max* représentent respectivement la similarité moyenne de x_i et de x_j avec l'ensemble des autres sommets ; si la similarité entre ces deux sommets est plus élevée que leur similarité moyenne respective, on estime que x_i et x_j sont "en moyenne" plutôt similaires. Le critère (1) permet ainsi d'éviter le recours à un seuil arbitraire.

Définition 2.2 *Soit $G_S(X, V)$ le graphe des similarités sur l'ensemble des objets X. Un* **pôle** *P_k est un sous-ensemble de X tel que le sous-graphe $G_S(P_k, V(P_k))$ est complet[1], où $V(P_k)$ représente l'ensemble des arêtes de V pour lesquelles les sommets sont dans P_k.*

Soient :	$X = \{x_1, \ldots, x_n\}$ l'ensemble des objets S la matrice de similarités sur $X \times X$
Initialisation :	Construire le graphe de similarités $G_S(X, V)$
Etape 1 :	Construire l'ensemble $\mathcal{P}$ des pôles $\{P_1, \ldots, P_l\}$ avec $\forall i \in \{1, \ldots, l\}$ $P_i \subseteq X$
Etape 2 :	Construire la matrice U des appartenances avec $u(P_i, x_j) = \frac{1}{\lvert P_i \rvert} \sum_{x_k \in P_i} s(x_j, x_k)$
Etape 3 :	Pour chaque $x_j \in X$, `affecter`(x_j,$\mathcal{P}$)
Etape 4 :	Soit $\mathcal{C}$ l'ensemble des groupes $\{C_1, \ldots, C_l\}$ tels que : $C_i = \{x_j \in X \lvert x_j$ est affecté à $P_i\}$ Construire l'arbre hiérarchique sur $\mathcal{C}$

TAB. 1 – PoBOC : algorithme de soft-clustering.

Le tableau TAB. 1 présente l'algorithme PoBOC. On peut noter que la fonction d'appartenance d'un objet à un pôle ($u(P_i, x_j)$), définie à l'étape 2, n'est pas une

[1] Un sous-graphe complet est aussi appelé "clique".

fonction probabiliste[2]. L'appartenance d'un objet à un pôle est donnée par la moyenne des similarités de cet objet avec chacun des objets du pôle considéré. Un "outlier" peut avoir des degrés d'appartenance faibles pour chacun des pôles.

Les heuristiques de construction des pôles, d'affectation multiple et de hiérarchisation des clusters sont présentées en section 2.2.

2.2 Heuristiques pour PoBOC

Heuristique de construction des pôles
La notion de *core* (noyau) présentée dans [Ben-Dor *et al.*, 1999] et de *foyer* (centroïde ou médoïde) utilisé en fuzzy-clustering sont assez proches de la notion de pôle que nous définissons ici. La construction des pôles consiste à rechercher un ensemble de cliques dans le graphe de similarités. La recherche d'une clique de taille maximale dans un graphe étant un problème NP-complet, on utilisera l'heuristique d'approximation *"Best in"* [Bomze *et al.*, 1999]. Cette heuristique procède, à partir d'un sommet de départ, par ajouts successifs du voisin le plus proche jusqu'à l'absence de voisins communs.

La construction de l'ensemble des pôles est donc une méthode itérative comportant deux tâches : le choix d'un sommet de départ et la construction d'une clique autour de ce sommet.

Le premier sommet x^1 est celui dont la similarité moyenne avec l'ensemble des objets est la plus faible, à condition qu'il dispose d'au moins un voisin dans le graphe. Soit $G_S(X, V)$ le graphe de similarités :

$$x^1 = Argmin_{x_i \in E} \frac{1}{|X|} \sum_{x_j \in X} s(x_i, x_j) \tag{2}$$

où E est l'ensemble des sommets tels que `degre(x)>0`[3] dans G.

Les sommets suivants $\{x^2, \ldots, x^l\}$ sont choisis de manière à s' "éloigner" au maximum des pôles déjà construits jusqu'à ce que cet éloignement soit jugé trop faible. Ainsi, pour une matrice de similarités S normalisée sur $[-1, 1]$:

$$x^k = Argmin_{x_i \in E} \frac{1}{k-1} \sum_{m=1,\ldots,k-1} \frac{1}{|P_m|} \sum_{x_j \in P_m} s(x_i, x_j) \tag{3}$$

à condition que la double somme calculée en (3) soit négative. Dans le cas contraire, on estimera qu'il n'existe plus de sommets suffisamment éloignés des pôles déjà construits. C'est ce critère qui détermine le nombre l de pôles (et donc de groupes terminaux).

Méthode d'affectation "soft"
C'est par cette étape d'affectation, ou plutôt "multi-affectation" des objets aux pôles que PoBOC se place dans une approche de "soft-clustering". Si beaucoup d'études abordent la nécéssité de pouvoir affecter un même objet à plusieurs groupes, peu d'entre elles proposent une méthode d'affectation efficace. Le plus souvent, l'utilisation d'un seuil (plus ou moins arbitraire) permet d'obtenir un tel résultat à partir de la matrice

[2]La somme des appartenances n'est pas égale à 1.
[3]Dans un graphe, le degré est égal au nombre de voisins du sommet.

d'appartenance floue obtenue par une méthode de *fuzzy clustering* [Kearns *et al.*, 1997]. Le recours à un seuil peut parfois suffire. Cependant, dans notre situation, l'organisation des données en groupes non-disjoints n'est qu'une étape dans le processus d'apprentissage, nous avons besoin pour cela d'un critère d'affectation général. Nous proposons alors la procédure suivante :

Définition 2.3 *Soient $X = \{x_1, \ldots, x_n\}$ l'ensemble des objets, $\mathcal{P} = \{P_1, \ldots, P_l\}$ l'ensemble des pôles et U la matrice d'appartenance sur $\mathcal{P} \times X$. Pour un objet x_j donné on note $P_{j,1}$ le pôle dont x_j est le plus "proche" ($P_{j,1} = Argmin_{P_i \in \mathcal{P}}\ u(P_i, x_j)$), $P_{j,2}$ le deuxième plus "proche" et ainsi de suite jusqu'à $P_{j,l}$, ce dernier étant alors le pôle le plus éloigné de x_j.*
`AFFECTER`*(x_j,$P_{j,k}$) si et seulement si l'une des trois propriétés suivantes est vraie :*

i) $k=1$
ii) $1 < k < l\ ,\ u(P_{j,k}, x_j) \geq \frac{u(P_{j,k-1},x_j)+u(P_{j,k+1},x_j)}{2}\ et\ \forall k' < k\ ,\ u(P_{j,k'}, x_j) = 1$
iii) $k = l\ et\ u(P_{j,k}, x_j) \geq \frac{u(P_{j,k-1},x_j)}{2}$

La propriété *i)* permet d'affecter chaque instance au moins au pôle qui lui est le plus proche. Les deux propriétés suivantes conduisent à affecter un objet à un autre pôle relativement à ses valeurs d'appartenance aux pôles immédiatement plus proche et plus éloigné. Le critère d'affectation est alors "universel" au sens où il dépend de la place de l'objet dans l'environnement des pôles.

Organisation hiérarchique des groupes
L'organisation hiérarchique a pour but de proposer une représentation à plusieurs niveaux de précisions. Nous verrons dans les applications proposées par la suite, que ce type d'organisation permet de mieux définir les relations existant entre les concepts terminaux obtenus et parfois de diminuer le nombre de groupes tout en conservant une organisation convenable pour la tâche d'apprentissage à venir.

Le principe de hiérarchisation revient à appliquer l'algorithme agglomératif hiérarchique du simple lien [Jain *et al.*, 1999] à partir des groupes déjà constitués $\mathcal{C} = \{C_1, \ldots, C_l\}$ où $C_i = \{x_j$ affectés à $P_i\}$. La matrice de similarités S étant normalisée sur $[-1, 1]$, on a $\forall x_i \in X\ s(x_i, x_i) = 1$ et on définit la similarité entre deux groupes non-disjoints par :

$$sim(C_k, C_m) = \frac{1}{|C_k|.|C_m|} \sum_{x_i \in C_k} \sum_{x_j \in C_m} s(x_i, x_j) \tag{4}$$

L'arbre hiérarchique est alors construit par fusions successives des deux plus proches groupes jusqu'à obtenir un seul groupe contenant tous les objets ; les feuilles de l'arbre (initialisation) étant constituées par les groupes résultant de l'étape d'affectation de PoBOC.

2.3 Discussion sur PoBOC

Une première remarque porte sur la complexité de l'algorithme, ce qui constitue souvent un argument en défaveur du soft-clustering ; en effet l'espace des possibilités

est beaucoup plus important dans le cas “soft” que pour les algorithmes de hard-clustering. Cependant, le clustering ne consiste pas, même dans le cas “hard” à évaluer toutes les possibilités. La complexité de l'algorithme PoBOC est bornée par $o(k.n^2)$, correspondant à l'étape de hiérarchisation (la plus coûteuse) où k est le nombre de clusters générés et n le nombre d'objets. De ce point de vue, PoBOC se situe entre les méthodes performantes telles que c-means ou fuzzy-c-means (linéaires sur le nombre d'objets) et d'autres plus coûteuses comme par exemple les approches agglomératives hiérarchiques du lien complet ou du lien moyen (en $o(n^2.log\ n)$).

D'autre part, si dans le cas “objet” les approches floues définissent des centroïdes, le passage au cas “relationnel” traduit les foyers en terme de médoïdes induisant alors une imprécision quant à la représentativité du foyer. La définition de pôles dans PoBOC, permet de limiter l'imprécision en considérant la similarité moyenne avec plusieurs objets. De même, le fait de ne pas fixer le nombre de groupes initialement et d'obtenir des résultats ne dépendant d'aucune initialisation aléatoire sont autant d'arguments conduisant à préférer PoBOC à d'autres approches telles que c-means et ses variantes.

3 Organisation des données pour l'apprentissage de règles de classification

De nombreux travaux existent en apprentissage supervisé pour apprendre, à partir d'un ensemble de données (ou instances) d'entraînement étiquetées, un classifieur permettant de déterminer correctement l'étiquette de nouvelles instances. On distingue deux principales approches : les méthodes basées sur les instances (Instance-Based Learning - IBL) et celles basées sur les règles (Rule-Based Learning - RBL). Dans le premier cas, il s'agit le plus souvent de définir sur les données une métrique appropriée, puis de déterminer l'étiquette d'une instance relativement à sa position dans l'ensemble des instances d'entraînement via cette métrique. Dans le cas des approches basées sur les règles, on cherche plutôt à construire une structure (arbre de décision, ensemble de règles...) basée sur les attributs décrivant les données et partitionnant de façon floue ou non l'espace des hypothèses.

L'hypothèse que nous formulons pour cette application de PoBOC est triple. D'une part nous cherchons à démontrer que l'organisation des données peut aider à construire des règles de meilleure qualité et ainsi améliorer la performance des approches RBL. D'autre part nous voulons confirmer l'idée selon laquelle les techniques d'organisation autorisant les recouvrements entre classes (soft-clustering) permettent une meilleure représentation des données que les techniques de regroupement dur. Enfin, parmi ces algorithmes de “soft-clustering”, nous voulons évaluer les performances de PoBOC.

3.1 Principe de la méthode

Les approches RBL dites “gloutonnes” construisent successivement des règles en se focalisant sur l'ensemble des instances non encore couvertes. Ces règles sont elles-mêmes construites par ajouts successifs de littéraux du type $attribut = valeur$ ou $attribut > valeur$ tels que chaque littéral couvre le plus possible d'instances positives et rejette le plus possible d'instances négatives. La conjonction de ces littéraux forme

le corps de la règle, la classe en constituant la tête. Ainsi chaque classe est caractérisée par la disjonction des règles apprises pour cette classe.

Soient : E l'ensemble des instances d'entraînement
$C_1, \ldots, C_k$ les classes à apprendre (en extension)

Initialisation : Construire une matrice S de similarités sur $E \times E$
$\mathcal{R} = \emptyset$

Pour chaque classe C_i :

1. $N = E \setminus C_i$ *(instances négatives)*
2. $F = \emptyset$ et `decomposer`(C_i,S) tel que $C_i = C_{i,1} \cup \ldots \cup C_{i,k}$
3. pour chaque cluster $C_{i,j}$:
   Si `construire_regle`($C_{i,j}$,N) abouti à une règle $R_{i,j}$
   alors $\mathcal{R} \leftarrow \mathcal{R} \cup R_{i,j}$
   Sinon $F \leftarrow F \cup C_{i,j}$
4. Si $F \neq \emptyset$ alors $C_i = F$ et GoTo 2

Retourner : L'ensemble $\mathcal{R}$ des règles apprises.

TAB. 2 – Apprentissage de règles par décomposition des classes.

Nous proposons dans TAB. 2, un algorithme général d'apprentissage de règles par décomposition de classes. Cet algorithme sera alors couplé avec différents algorithmes de clustering, dont PoBOC. On note en premier lieu la nécessité de recourir à une mesure de similarité traitant aussi bien des données décrites par des attributs numériques que symboliques. Pour cela [Martin et Moal, 2001] proposent une mesure basée sur la définition d'un nouveau language de description engendré de manière aléatoire à partir des attributs initiaux. Le nombre de termes constituant ce language détermine la précision de la mesure.

Cette mesure de similarité permet ensuite d'organiser chaque classe en sous-classes. Chaque sous-classe est soumise au test d'existence d'une règle couvrant toutes les instances de cette sous-classe et aucune instance négative[4]. Dans le cas où une telle règle n'existe pas, l'union des sous-classes non couvertes par une règles est de nouveau décomposée. Finalement, chaque groupe ainsi construit est généralisé par une règle à laquelle est associée un score de fiabilité afin de gérer les conflits (cf. [Ali et Pazzani, 1993]).

[4]On autorisera les règles à couvrir des exemples positifs d'autres sous-classes, de même qu'on sera parfois amené à tolérer le fait que la règle couvre quelques exemples négatifs (tolérance au bruit).

3.2 Evaluations et discussion

La méthode est évaluée sur quelques jeux de données classiques de l'UCI repository [Merz et Murphy, 1998]. Lorsqu'un ensemble test n'est pas fourni, le taux de bonne classification est calculé sur la moyenne de 10 validations croisées. Les méthodes de classification proposées sont évaluées à chaque fois sur les mêmes échantillons test, ceci permet de comparer les performances de chaque classifieur dans des conditions identiques.

Le tableau Tab. 3 présente les résultats obtenus en utilisant plusieurs algorithmes de décomposition : $CLINK$ (algorithme agglomératif hiérarchique du lien complet), $CMED$ (algorithme de partitionnement des c-médoïdes[5]), PoBOC et FCMdd (Fuzzy-c-Médoïdes avec la procédure d'affectation présentée dans la section 2.2).

DOMAINES	CLINK	CMED	FFCMdd	PoBOC
AUDIOLOGY	88.9 (1)	79.6 (4)	80.1 (3)	81.2 (2)
IRIS	95.3 (3)	95.3 (3)	95.7 (1)	95.7 (1)
SOYBEAN	73.8 (3)	71.4 (4)	74.0 (2)	78.2 (1)
WINE	93.3 (4)	94.1 (3)	95.3 (1)	94.9 (2)
ZOOLOGY	90.9 (1)	90.3 (2)	90.2 (3)	90.2 (3)
Position moyenne	2.4	3.2	2.0	**1.8**

Tab. 3 – Comparaison des algorithmes de clustering (% de bonne classification et position du classifieur).

Ce premier tableau met en évidence la supériorité (en terme de classement) de PoBOC et plus généralement des approches de regroupement avec recouvrement (soft-clustering). Cependant on notera la variabilité des résultats d'un domaine à un autre et notamment le taux de classification nettement meilleur de CLINK pour le domaine “Audiology” et de PoBOC sur “Soybean”. Cette première étude est confirmée par le tableau Tab. 4 qui permet d'appréhender la difficulté avec laquelle chaque méthode a abouti à un ensemble de sous-classes généralisables par une règle.

Enfin, le tableau Tab. 5 présente quelques éléments de comparaison entre pFOIL (apprentissage de règles sans organisation des données) et Clust-PoBOC (apprentissage de règles via clustering par PoBOC) ainsi qu'avec d'autres approches de classification bien connues : classification par arbres de décision avec C4.5 [Quinlan, 1986] et par le plus proche voisin (1-*Nearest Neighbor*). Ce résultat montre clairement l'intérêt d'organiser les données pour l'apprentissage.

Algorithmes de clustering	CLINK	CMED	FFCMdd	PoBOC
Nb. moyen d'appels de `decompose`	4.25	5.3	5.35	1.95
Nb. moyen de règles	7.5	7.8	8.5	5.8

Tab. 4 – Nombres moyens par classe (sur les classes 2 et 3 d'*Iris*) de règles générées et d'appels de la procedure de décomposition. Moyenne sur une validation croisée.

[5]Variante de c-means dans le cas de données relationnelles.

DOMAINES	Clust-PoBOC	pFOIL	C4.5+élaguage	1-NN
AUDIOLOGY	81.2 (1)	60.3 (4)	70.3 (3)	77.7 (2)
IRIS	95.7 (1)	93.8 (4)	95.5 (2)	94.7 (3)
SOYBEAN	78.2 (2)	71.3 (4)	86.7 (1)	75.3 (3)
WINE	94.9 (1)	91.7 (4)	94.1 (3)	94.6 (2)
ZOOLOGY	90.2 (4)	90.7 (3)	91.2 (2)	94.9 (1)
Position moyenne	**1.8**	3.8	2.2	2.2

TAB. 5 – Amélioration du classifieur en organisant les données (% de bonne classification et position du classifieur).

4 Organisation de termes pour le traitement des données textuelles

4.1 Problématique générale des données textuelles

Les approches d'organisation par soft-clustering prennent tout leur sens dans l'application aux données issues du langage naturel. En effet, les mots et textes sont considérés comme des informations complexes et difficiles à caractériser et à représenter fidèlement. Losqu'il s'agit de regrouper des mots en classes sémantiques, il est parfois difficile et souvent réducteur de choisir pour un mot donné une et une seule classe (c'est encore plus vrai dans le cas des mots polysémiques). Finalement, dans ce type d'application, la difficulté est double puisqu'il s'agit de deux traitements inter-dépendants : la définition d'une bonne mesure de similarité (entre mots ou documents) et le choix d'une méthode de regroupement adaptée.

Nous nous intéressons ici à l'organisation de termes. Nous définissons un terme comme étant composé de un ou plusieurs mots dont l'association fait référence à un sens. Typiquement un "mot-clé" répond à cette définition. L'évaluation d'une telle organisation est difficile à quantifier, c'est pourquoi nous proposons de travailler sur une petite base de données afin de pouvoir représenter l'organisation des termes dans son ensemble et laisser l'expert (ici le lecteur) analyser le résultat. Toutefois, des classes de termes sont pré-définies par la source d'extraction des données (thématique des documents desquels les termes sont issus) permettant ainsi de disposer de quelques éléments de comparaisons.

4.2 Expérimentations et discussion

Nous travaillons sur les mots-clés d'articles scientifiques sur trois domaines de spécialités : l'Intelligence Artificielle, le Web et les Technologies de la Langue Naturelle. Nous avons extrait au hasard 38 mots-clés dans les conférences ou revues respectives : JSAI 1997[6], WWW 2002[7] et LREC 2000[8]. La mesure de similarité utilisée pour évaluer

[6] *Journal of Japanese Society for Artificial Intelligence*
[7] *Eleventh International World Wide Web Conference*
[8] *2nd International Conference on Language Resources and Evaluation*

la proximité sémantique entre les termes est basée sur les cooccurrences des termes sur Internet, cette mesure a été présentée dans [Clavier *et al.*, 2002]. Les termes sont alors regroupés de façon totalement non-supervisée. Cependant une évaluation est possible sur la pureté des classes obtenues comparativement aux classes d'origine des termes.

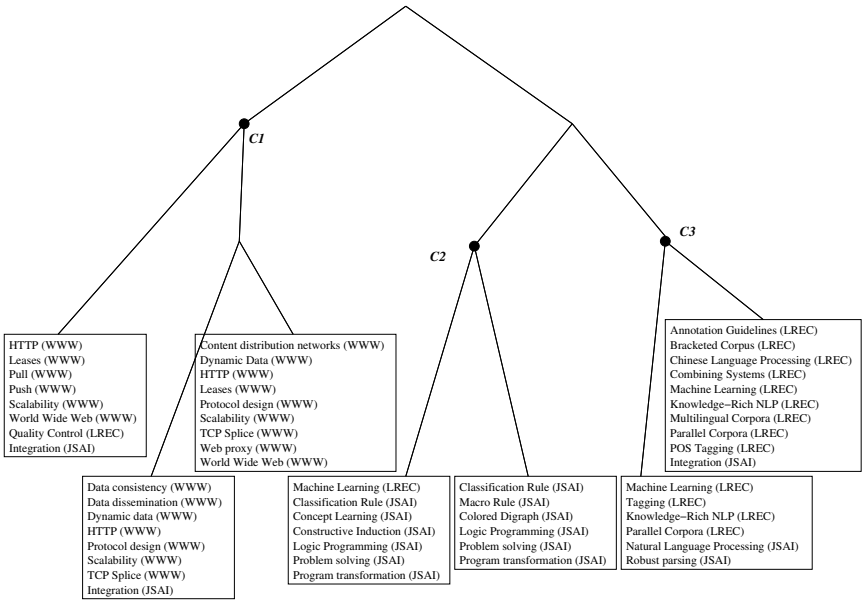

FIG. 1 – Organisation hiérarchique des termes par PoBOC

La figure FIG. 1 présente l'arbre hiérarchique relatif à l'organisation des 38 mots-clés, renvoyé par PoBOC. Les noeuds C_1, C_2 et C_3, en haut de l'arbre, correpsondent aux trois classes d'origine avec une puretée moyenne de 88%. En descendant dans cet arbre on trouve 7 sous-classes correspondant aux groupes constitués autour des pôles, la pureté moyenne de ces groupes est de 86%. Ces résultats quantitatifs sous-estiment la qualité de l'organisation si on considère les mots-clés, à priori mal placés, alors qu'en fait ils se retrouvent dans leur vraie classe sémantique, c'est le cas de *“Natural Language Processing”*, *“Robust parsing”* ou encore *“Machine Learning”*. PoBOC autorise le recouvrement entre les classes, ce qui permet de distinguer certains mots-clés généraux (*“Integration”*) et d'autres au centre d'un thème (*“HTTP”, “Scalability”, “Logic Programming”*). Enfin, PoBOC fournit un résultat unique et indépendant de toute initialisation, contrairement aux approches telles que *Fuzzy-c-medoid* qui, avec $c = 3$, donne un taux de pureté moyen des classes de 69% seulement, avec des variations entre 55% et 87%.

5 Conclusion

Nous avons présenté dans cette étude l'algorithme de regroupement PoBOC ayant la particularité de construire des groupes non-disjoints (“soft-clustering”). Cette ca-

ractéristique répond aux exigences des applications de plus en plus étudiées en apprentissage, notamment le traitement des données textuelles. Nous avons démontré que PoBOC conduit à améliorer la qualité de l'organisation des données sur deux domaines très différents : l'apprentissage de règles de classification et le regroupement de termes issus du langage naturel. Dans le premier cas nous avons pu mettre en évidence que l'organisation des données en groupes non-disjoints permet d'améliorer la qualité des règles apprises ; l'application proposée sur les données textuelles est venue conforter l'intérêt d'utiliser PoBOC afin de représenter plus fidèlement la structure intrinsèque des données.

L'utilisation de cet algorithme sur des données textuelles plus volumineuses se placera dans la continuité de cette étude. En effet, certains travaux présentent l'intérêt de regrouper les mots contenus dans les documents afin de mieux classer ces documents. Nous pouvons penser que cette classification sera d'autant meilleure que les groupes de termes seront représentatifs de concepts sémantiques. En ce sens PoBOC peut constituer un outil clé dans la construction de telles classes sémantiques.

Références

[Agrawal *et al.*, 1992] R. Agrawal, S. Ghosh, T. Imielinski, B. Iyer, et A. Swami. An interval classifier for database mining applications. In Li-Yan Yuan, editor, *Proceedings of the 18th International Conference on Very Large Databases*, pages 560–573, San Francisco, U.S.A., 1992. Morgan Kaufmann Publishers.

[Ali et Pazzani, 1993] K. M. Ali et M. J. Pazzani. HYDRA : A noise-tolerant relational concept learning algorithm. In R. Bajcsy, editor, *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1064–1071. Morgan Kaufmann, 1993.

[Baker et McCallum, 1998] L. D. Baker et A. K. McCallum. Distributional clustering of words for text classification. In W. Bruce Croft, Alistair Moffat, Cornelis J. van Rijsbergen, Ross Wilkinson, et Justin Zobel, editors, *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 96–103, Melbourne, AU, 1998. ACM Press, New York, US.

[Ben-Dor *et al.*, 1999] A. Ben-Dor, R. Shamir, et Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3/4) :281–297, 1999.

[Bomze *et al.*, 1999] I. Bomze, M. Budinich, P. Pardalos, et M. Pelillo. The maximum clique problem. In D.-Z. Du et P. M. Pardalos, editors, *Handbook of Combinatorial Optimization*, volume 4. Kluwer Academic Publishers, Boston, MA, 1999.

[Clavier *et al.*, 2002] V. Clavier, G. Cleuziou, et L. Martin. Organisation conceptuelle de mots pour la recherche d'information sur le web. In *Conférence d'Apprentissage CAp'2002*, pages 220–235. PUG Presses Universitaires de Grenoble ISBN 2 7061 1092 9, 2002.

[Diday, 1986] E. Diday. Une représentation visuelle des classes empiétantes : Les pyramides. In *Rairo : Analyse des Données (vol. 52)*, pages 475–526, 1986.

[Jain *et al.*, 1999] A. K. Jain, M. N. Murty, et P. J. Flynn. Data clustering : a review. *ACM Computing Surveys*, 31(3) :264–323, 1999.

[Jain *et al.*, 2000] A. K. Jain, R. P. W. Duin, et Jianchang Mao. Statistical pattern recognition : A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1) :4–37, 2000.

[Kearns *et al.*, 1997] M. Kearns, Y. Mansour, et A. Y. Ng. An information-theoretic analysis of hard and soft assignment methods for clustering. In *Proceedings of Uncertainty in Artificial Intelligence. AAAI*, pages 282–293, 1997.

[Lin et Kondadadi, 2001] K. I. Lin et R. Kondadadi. A word-based soft clustering algorithm for documents. In *Proceedings of 16th International Conference on Computers and Their Applications*, 2001.

[MacQueen, 1967] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical statistics and probability*, volume 1, pages 281–297. University of California Press, 1967.

[Martin et Moal, 2001] L. Martin et F. Moal. A language-based similarity measure. In *Machine Learning : ECML 2001, 12th European Conference on Machine Learning, Freiburg, Germany, September 5-7, 2001, Proceedings*, volume 2167 of *Lecture Notes in Artificial Intelligence*, pages 336–347. Springer, 2001.

[Merz et Murphy, 1998] C.J. Merz et P.M. Murphy. Uci repository of machine learning databases. 1998.

[Pham et Prince, 1998] D. Pham et J. Prince. An adaptive fuzzy c-means algorithm for image segmentation in the presence of intensity inhomogeneities. In *Proceedings SPIE Medical Imaging 1998 : Image Processing*, volume 3338, pages 555–563, 1998.

[Quinlan, 1986] J.R. Quinlan. Induction of decision trees. In *Machine Learning*, pages 81–106, 1986.

[Sneath et Sokal, 1973] P. H. A. Sneath et R. R. Sokal. Numerical taxonomy - the principles and practice of numerical classification. 1973.

Summary

We describe the PoBOC algorithm (Pole-Based Overlapping Clustering) which builds a set of non disjoint clusters (“soft-clusters”) hierarchically organized, from the similarity matrix over the considered data. The approach is tested on two learning tasks : rules learning for a classification task and the organization of more complex and weakly structured data as textual ones.

The evaluation of clustering methods is a difficult process. The two applications proposed allow to validate the organization method with a quantitative point of view (classification) and with a “human” evaluation (textual data clustering). We show the significance of a previous organization of the data before the final learning task and we conclude on the interest of PoBOC comparatively with other unsupervised approaches.