De l'utilisation de la proportion analogique en apprentissage artificiel

Laurent Miclet, Sabri Bayoudh, Arnaud Delhay, Harold Mouchère

IRISA et ENSSAT
6, rue de Kérampont, BP 80518, 22305 Lannion Cedex
{miclet,bayoudh,delhay,mouchere}@irisa.fr

Résumé : Cet article s'intéresse à la *proportion analogique*, une forme simple du raisonnement par analogie, et décrit son utilisation en apprentissage artificiel. Nous nous attachons plus particulièrement à définir une nouvelle notion, la *dissimilarité analogique* et à l'appliquer à des séquences. Après avoir défini la proportion analogique, la dissimilarité analogique et la résolution approchée d'équations analogiques, nous décrivons deux algorithmes qui rendent opérationnels ces notions pour des objets numériques ou symboliques et pour des séquences de ces objets. Nous montrons ensuite leur efficacité au travers de deux cas pratiques : le premier est l'apprentissage d'une règle de classification pour des objets décrits par des attributs nominaux ; le second montre comment la génération de nouveaux exemples (par résolution approchée d'équations analogiques) peut aider un système de reconnaissance de caractères manuscrits à s'adapter très rapidement à un nouveau scripteur. Une discussion plus générale sur l'apport du raisonnement par analogie pour l'apprentissage artificiel termine cet article.

Mots-clés : Apprentissage, classification, analogie.

1 Introduction

Cet article se propose d'appliquer certaines notions du raisonnement par analogie dans le domaine de l'apprentissage supervisé non paramétrique. En apprentissage supervisé, on dispose d'un ensemble d'apprentissage $\mathcal{S}$ composé d'objets, chacun étant associé à sa supervision (une étiquette de classe, par exemple). La question est de savoir à quelle classe on doit affecter un nouvel objet, à partir de la seule connaissance de l'ensemble d'apprentissage (CORNUÉJOLS & MICLET (2002)).

Le principe de l'apprentissage non paramétrique est de ne faire aucune hypothèse sur la distribution statistique des classes. La technique la plus simple de ce domaine est celle de l'apprentissage par *plus proche voisin* : quand arrive un objet extérieur à $\mathcal{S}$, on lui attribue la classe de l'élément le plus ressemblant dans $\mathcal{S}$.

Dans la même famille, la technique de l'apprentissage par analogie fait appel à un argument plus sophistiqué. Donnons-en un exemple intuitif sur des objets qui sont représentés par des séquences de lettres. Soit la séquence `cherchera`, dont on veut déterminer la classe ; supposons que dans l'ensemble d'apprentissage se trouvent les

trois séquences : `recommencer`, `commencera`, `rechercher`, avec respectivement pour classes *infinitif*, *futur*, *infinitif*. On attribuera à `cherchera` la classe *futur*, par un raisonnement qui s'énonce comme ceci. Puisque, dans l'univers des séquences

`recommencer` *est à* `commencera` *comme* `rechercher` *est à* `cherchera`

la supervision de `cherchera` est donc la solution de l'équation sur les classes :

infinitif est à *futur* comme *infinitif* est à x

d'où : $x =$ *futur*.

Ce petit exemple fait apparaître deux notions de base : d'abord celle de *proportion analogique entre quatre objets* (des séquences de lettres dans l'exemple ci-dessus) qui s'exprime sous la forme "A *est à* B *comme* C *est à* D". Il faut bien sûr donner une signification à « comme » et à « est à ». En effet, dans un autre exemple, la proportion

"`jument` *est à* `poulain` *comme* `vache` *est à* `veau`"

porte seulement sur la sémantique des mots, pas sur leur morphologie.

Nous verrons que cette notion peut s'étendre à une proportion approchée, que l'on pourrait exprimer par "A *est à* B *à peu près comme* C *est à* D". Par exemple :

"`jument` *est à* `poulain` *à peu près comme* `vache` *est à* `bufflon`"
"`recommencer` *est à* `commencera` *à peu près comme* `rechercher` *est à* `chercherai`"

Nous donnerons une quantification et une méthode de calcul du terme « à peu près ».

L'autre notion que cet exemple introduit est celle d'*équation analogique* : si l'on connaît trois termes d'une proportion analogique, peut-on calculer le quatrième ? Et peut-on calculer tous les termes qui sont « à peu près » en proportion analogique avec les trois premiers ? Nous donnerons des algorithmes pour résoudre ce problème et nous montrerons en quoi ils sont utiles pour l'apprentissage supervisé.

D'une manière générale, la proportion analogique est un cas particulier du *raisonnement par analogie* qui a été longuement décrit et étudié depuis les philosophes grecs ; ses applications récentes intéressent en particulier les sciences cognitives (Holyoak (2005)), la linguistique et l'intelligence artificielle. LEPAGE (2003) donne une histoire encyclopédique de ce concept et de ses applications à la science du raisonnement et à la linguistique. La restriction à la proportion analogique, en particulier pour les séquences, a été étudiée d'un point de vue méthodologique et algorithmique en particulier dans MITCHELL (1993), HOFSTADTER & the Fluid Analogies Research Group (1994), DASTANI *et al.* (2003)), SCHMID *et al.* (2003), YVON *et al.* (2004). Un autre domaine de l'Intelligence Artificielle est naturellement relié au raisonnement par analogie : celui du raisonnement à base de cas (CBR). Dans le livre de référence AAMODT & PLAZA (1994), ces deux notions sont confondues. Nous reviendrons dans la conclusion sur la distinction entre CBR et apprentissage par proportion analogique et sur la légitimité de ce dernier.

2 Proportion analogique et équations analogiques

2.1 Les axiomes de la proportion analogique

Il n'y a pas de définition générale d'une proportion analogique "A *est à* B *comme* C *est à* D" entre quatre objets pris dans un ensemble X, les relations "est à" et "comme" dépendant de la nature de X. Toutefois, d'après la signification usuelle du mot "analogie" en philosophie et en linguistique, trois axiomes de bases sont généralement requis (LEPAGE & ANDO (1996)) :

Definition 1 (Proportion analogique.)
Une proportion analogique sur X est une relation sur X^4. Quand $(A, B, C, D) \in \mathcal{A}$, les quatre éléments A, B, C et D sont dits en proportion analogique*, ce qui s'écrit $A : B :: C : D$ et se lit "A est à B comme C est à D". Pour chaque 4-uplet en proportion analogique, les trois axiomes suivants sont requis :*

S*ymétrie de la relation "comme" :* $A : B :: C : D \Leftrightarrow C : D :: A : B$
E*change des médians :* $A : B :: C : D \Leftrightarrow A : C :: B : D$
D*éterminisme :* $A : A :: B : x \Rightarrow x = B$

2.2 Equations analogiques

Résoudre une équation analogique consiste à trouver le quatrième terme d'une proportion analogique, les trois premiers étant connus.

Definition 2 (Equation analogique.)
D est une solution à l'équation analogique $A : B :: C : x$ ssi $A : B :: C : D$.

Selon la nature des objets et la définition des relations, une équation analogique peut ne pas avoir de solution, avoir une solution unique ou avoir plusieurs solutions. Nous étudions au paragraphe 3.3 la résolution des équations analogiques dans différents ensembles d'objets et dans des structures séquencielles de ces objets, en proposant une manière leur trouver des solutions approchées si elles n'ont pas de solution.

2.3 Proportions analogiques sur $\mathbb{R}^n$ et sur $\{0, 1\}^n$

Quand on se place dans $\mathbb{R}^n$, il est simple de remarquer que quatre objets sont en proportion analogique quand ils forment un parallélogramme, ce qui peut s'écrire :

$$a : b :: c : d \Leftrightarrow \overrightarrow{Oa} + \overrightarrow{Od} = \overrightarrow{Ob} + \overrightarrow{Oc} \Leftrightarrow \overrightarrow{ab} = \overrightarrow{cd}$$

Il est simple aussi de remarquer que quatre objets binaires vérifient les axiomes de la proportion analogique soit quand ils ont tous les quatre la même valeur 0 ou 1, soit quand deux valent 1 et deux valent 0 (sauf dans les cas (0,1,1,0) et (1,0,0,1)). Il y a donc au total 6 quadruplets binaires en analogie parmi les 16 possibles. Quand les objets sont des éléments de $\{0, 1\}^n$, il suffit que chaque quadruplet de coordonnées vérifie l'une des 6 proportions analogiques précédentes pour que les objets vérifient les axiomes de la proportion analogique. Ces définitions, ainsi que celles dans d'autres types d'ensembles comme les groupes cycliques, sont détaillées dans MICLET & DELHAY (2005).

2.4 Proportions analogiques sur les séquences

Nous utilisons maintenant les notions usuelles de la théorie des langages : alphabet Σ, mot (ou séquence) sur $\Sigma^\star$, longueur d'un mot, mot vide ϵ, facteur, sous-séquence. Par exemple, sur l'alphabet $\Sigma = \{a, b\}$, la séquence $aabbaa$ est un élément de $\Sigma^\star$ de longueur $|aabbaa| = 6$, $bbaa$ en est un facteur et aba en est une sous-séquence.

Nous ajoutons une nouvelle lettre à Σ, que nous notons $\backsim$, pour obtenir un alphabet augmenté Σ'. Son interprétation est celle d'un symbole "vide" nécessaire pour les sections qui suivent.

Definition 3 (Equivalence sémantique.)
Soit x une séquence de $\Sigma^\star$ et y une séquence sur $\Sigma'^\star$. Les séquences x et y sont sémantiquement équivalentes *si la sous-séquence de y composée des lettres de Σ est x. Nous notons cette relation par $\equiv$. Par exemple : $ab \backsim a \backsim a \equiv abaa$.*

Un alignement est une correspondance lettre à lettre entre quatre séquences, dans lesquelles des lettres $\backsim$ peuvent être insérées de façon à ce qu'elles prennent la même longueur. La correspondance $(\backsim, \backsim, \backsim, \backsim)$ n'est pas permise.

Definition 4 (Alignement entre quatre séquences.)
Un alignement *entre quatre sequences $u, v, w, x \in \Sigma^*$, est un mot z sur l'alphabet $(\Sigma \cup \{\backsim\})^4 \backslash \{(\backsim, \backsim, \backsim, \backsim)\}$ dont la projection sur la première, la seconde, la troisième et la quatrième composante sont sémantiquement équivalentes à u, v, w and x.*

Nous supposons qu'il existe une relation de proportion analogique dans Σ', c'est-à-dire que pour chaque quadruplet a, b, c, d dans Σ', la relation $a : b :: c : d$ vaut soit $VRAI$ soit $FAUX$. Nous proposons maintenant de définir la proportion analogique entre quatre séquences d'objets en utilisant à la fois la proportion analogique entre les objets qui les composent et l'alignement entre les quatre séquences.

Definition 5 (Proportion analogique entre séquences.)
Soit u, v, w and x quatre séquences de Σ, sur lequel existe une proportion analogique. Nous disons que u, v, w et x sont en proportion analogique *s'il existe quatre séquences u', v', w' and x' de même longueur dans Σ', avec les propriétés suivantes :*

1. *$u' \equiv u$, $v' \equiv v$, $w' \equiv w$ et $x' \equiv x$.*
2. *$\forall i \in [1, n]$ sont des proportions analogiques $u'_i : v'_i :: w'_i : x'_i$ dans Σ'.*

Par exemple, soit $\Sigma' = \{a, b, \alpha, \beta, B, C, \backsim\}$ avec les proportions analogiques $a : b :: A : B$, $a : \alpha :: b : \beta$ and $A : \alpha :: B : \beta$. L'alignement suivant entre les quatres séquences aBA, αbBA, ba et βba est une proportion analogique sur $\Sigma^\star$:

$$\begin{array}{cccc} a & \backsim & B & A \\ \alpha & b & B & A \\ b & \backsim & a & \backsim \\ \beta & b & a & \backsim \end{array}$$

Nous traitons au paragraphe 3.3 de la résolution des équations analogiques dans les séquences, à partir de cette définition.

3 Dissimilarité analogique

Nous proposons dans cette section de généraliser la relation de proportion analogique entre quatre objets ou quatre séquences d'objets en introduisant la notion de *dissimilarité analogique* (DA). Pour des raisons méthodologiques et opérationnelles, il est souhaitable que la DA soit à la distance ce que la proportion analogique est à l'égalité, c'est à dire qu'elle vérifie les propriétés suivantes :

Cohérence analogique. $DA(u,v,w,x) = 0 \Leftrightarrow u : v :: w : x$

Symétrie de "comme". $DA(u,v,w,x) = DA(w,x,u,v)$

Echange des médians. $DA(u,v,w,x) = DA(u,w,v,x)$

Inégalité triangulaire. $DA(u,v,z,t) \leq DA(u,v,w,x) + DA(w,x,z,t)$

Asymétrie de "est à". En général, $DA(u,v,w,x) \neq DA(v,u,w,x)$

3.1 Dissimilarité analogique entre objets de $\mathbb{R}^n$ et de $\{0,1\}^n$

Une DA dans $\mathbb{R}^n$ se définit naturellement par $DA(a,b,c,d) = \delta(d,e)$, où e est le quatrième sommet du parallélogramme construit sur a, b et c et où δ est une distance dans $\mathbb{R}^n$. Il est alors facile de vérifier que les propriétés précédentes sont vraies.

Pour quatre objets binaires, nous savons depuis le paragraphe 2.3 qu'il y a six cas de proportions analogiques exactes sur les 16 quadruplets binaires possibles. En affectant la valeur $DA(a,b,c,d) = 0$ à ces six quadruplets, la valeur 1 à tous ceux qui ont trois 0 et un 1 (ou le contraire) et la valeur 2 aux quadruplets $(0,1,1,0)$ et $(1,0,0,1)$, on démontre que les propriétés ci-dessus sont vérifiées.

Pour des objets de $\{0,1\}^n$, additionner les valeurs précédentes sur les coordonnées (puis diviser par n si l'on veut normaliser) produit également une dissimilarité analogique ayant les quatre propriétés désirées (MICLET & DELHAY (2005)). C'est cette définition que nous utiliserons dans la suite.

3.2 Dissimilarité analogique entre séquences : définition et calcul.

Nous pouvons maintenant définir la notion de DA sur des séquences d'objets pour lesquels une DA a été définie (et étendue pour prendre en compte le symbole $\backsim$).

Definition 6 (Dissimilarité analogique entre séquences.)
Le coût *d'un alignement entre quatre séquences est la somme des dissimilarités analogiques entre les quadruplets de lettres définis par cet alignement. La* dissimilarité analogique entre quatre séquences *est le coût de l'alignement le moins coûteux entre les quatre séquences.*

La définition précédente permet de calculer la dissimilarité analogique $DA(u,v,w,x)$ avec un algorithme de programmation dynamique (appelé SEQUANA4), qui progresse de façon synchrone dans les quatres séquences pour construire un alignement optimal. L'entrée de cet algorithme est un alphabet augmenté Σ' sur lequel une dissimilarité analogique $DA(a,b,c,d)$ a été définie. La sortie est la dissimilarité analogique entre

quatre séquences $\Sigma^\star$, c'est-à-dire $DA(u, v, w, x)$. Cet algorithme a une complexité en temps en $\mathcal{O}\big(|u|.|v|.|w|.|x|\big)$.

La validité de SEQUANA4 peut être montrée par récurrence, car il est construit sur le principe de la programmation dynamique. La DA calculée entre les séquences possède les propriétés suivantes : *cohérence avec l'analogie*, *symétrie du "comme"* et *l'échange des médians*. La propriété d'*inégalité triangulaire* n'est pas en général assurée.

3.3 Solutions approchées aux équations analogiques

Nous traitons dans ce paragraphe du calcul de la solution d'une équation analogique entre séquences, et ce d'une manière étendue : nous donnons en effet un algorithme capable de produire toutes les meilleures solutions à ce problème, c'est à dire de construire toutes les séquences en dissimilarité analogique minimale avec trois séquences données.

Definition 7 (Meilleure solution approchée à une équation analogique.)
Soit X un ensemble sur lequel est défini une analogie et une dissimilarité analogique DA. Soit $a : b :: c : x$ une équation analogique dans X. L'ensemble des meilleures solutions approchées à cette équation est donnée par :

$$\big\{y \ : \ \underset{y \in X}{ArgMin}\ DA(a, b, c, y)\big\}$$

Ce sont donc les objets $y \in X$ qui sont les plus proches d'être en proportion analogique avec a, b et c. Cette définition s'applique au cas de la solution exacte à une équation analogique quand on force la dissimilarité analogique à être nulle.

Nous pouvons facilement élargir ce concept et définir l'ensemble de k-meilleures solutions à une équation analogique $a : b :: c : x$. Informellement, c'est le sous-ensemble de k éléments de X qui ont une DA minimale quand ils sont associés en quatrième position de l'équation avec a, b et c.

Dans $\mathbb{R}^n$, il n'y a qu'une seule meilleure solution approchée à une équation analogique, qui peut-être explicitement calculée. Dans $\{0, 1\}^n$, l'approche naïve est d'examiner chaque élément de l'ensemble et de garder les y qui minimisent $DA(a, b, c, y)$. Comme dans ce cas DA possède la propriété d'inégalité triangulaire, nous avons montré dans Bayoudh *et al.* (2007) qu'un algorithme plus rapide pouvait être utilisé.

Le cas des séquences : l'algorithme SOLVANA

Nous décrivons maintenant un algorithme (appelé SOLVANA) qui trouve l'ensemble des meilleures solutions approchées à l'équation $a : b :: c : x$ quand les objets sont des séquences sur un alphabet étendu, sur lequel une DA a été définie.

Cet algorithme utilise la programmation dynamique pour construire un tableau à 3 dimensions. Quand cette construction est terminée, un retour en arrière est fait pour produire un graphe de toutes les meilleures solutions.

L'alignement de quatre séquences de longueurs différentes est réalisé en insérant optimalement des lettres $\backsim$ dans les trois séquences connues et en calculant en ligne la quatrième séquence, de telle façon que les quatre séquences aient finalement la même

longueur. En parallèle, on cumule les dissimilarités analogiques dans l'alphabet augmenté sur chaque étape de l'alignement des séquences. La complexité algorithmique de cet algorithme est en $O(m * p^3)$, où $m = Card(\Sigma')$ et p est la longueur moyenne des séquences.

Si l'on cherche non pas l'ensemble des solutions optimales, mais les k meilleures solutions pour k quelconque, cet algorithme n'est plus adapté. Il faut alors employer une technique du type *branch and bound*, comme l'algorithme $A^\star$ (en version de base avec une heuristique identiquement nulle ou de manière plus élaborée si l'on peut définir une fonction heuristique admissible).

4 Applications en apprentissage artificiel

4.1 Classification d'objets binaires et nominaux

Règle de classification par analogie

Soit $\mathcal{S} = \big\{\big(c_i, h(c_i)\big) \mid 1 \leq i \leq m\big\}$ l'ensemble d'apprentissage, avec $h(c_i)$ la classe de l'objet c_i. Les objets sont définis par des attributs binaires (les attributs nominaux sont binarisés par codage *one per value*). Soit x un objet n'appartenant pas à $\mathcal{S}$. Pour trouver la classe de x, on définit une règle d'apprentissage par analogie dépendant d'un entier k par les étapes suivantes :

1. Calculer la dissimilarité analogique entre x et touts les n triplets appartenant à $\mathcal{S}$ dont la résolution dans les classes a la forme suivante : "$\omega_i \ : \ \omega_i \ :: \ \omega_j \ : ?$", ce qui donne "$? = \omega_j$".
2. Ordonner ces n triplets par ordre croissant par rapport à leurs valeurs de DA, quand ils sont associés à x.
3. Si le k^{eme} triplet a la valeur p, alors soit k' le plus grand nombre tel que le triplet k'^{eme} a la même valeur p.
4. Résoudre les k' équations analogiques sur les classes. Choisir la classe gagnante qui comptabilise le plus grand nombre de votes parmi les k' résultats.

Pondération des attributs pour la règle de classification par analogie

Nous ne détaillons pas dans cette section la méthode de pondération (voir Bayoudh *et al.* (2007)) mais nous en donnons l'idée de base : puisque tous les attributs n'ont pas la même importance pour la classification, il faut donner une plus grande importance aux plus discriminants. Cependant, dans la classification par analogie, il y a plusieurs façons de conclure l'appartenance d'un élément à une certaine classe. C'est pourquoi nous avons défini un poids par classe pour chaque attribut.

Resultats

Nous avons appliqué la classification par proportion analogique ($WAPC$: Weighted Analogical Proportion Classifier) sur huit bases de données prise du UCI Repository de NEWMAN *et al.* (1998). Afin de mesurer l'efficacité du classificateur $WAPC$, nous l'avons comparé à six classificateurs standard. Les résultats obtenus sont donnés dans le tableau 1. Ils montrent la qualité et la polyvalence de la méthode $WAPC$ (problèmes multiclasses, données manquantes, etc.).

Methods	*MO.1*	*MO.2*	*MO.3*	*SP.*	*B.S*	*Br.*	*H.R*	*Mu.*
nombre d'attributs nominaux	7	7	7	22	4	9	4	22
nombre d'attributs binaires	15	15	15	22	4	9	4	22
nombre d'instances d'apprentissage	124	169	122	80	187	35	66	81
nombre d'instances de test	432	432	432	172	438	664	66	8043
nombre de classes	2	2	2	2	3	2	4	2
WAPC ($k = 100$)	**98%**	**100%**	**96%**	**79%**	**86%**	**96%**	**82%**	**98%**
APC ($k = 100$)	98%	100%	96%	58%	86%	91%	74%	97%
Decision Table	100%	64%	97%	65%	67%	86%	42%	99%
Id3	78%	65%	94%	71%	54%	mv	71%	mv
PART	93%	78%	98%	81%	76%	88%	82%	94%
Multi layer Perceptron	100%	100%	94%	73%	89%	96%	77%	96%
LMT	94%	76%	97%	77%	89%	88%	83%	94%
IB1	79%	74%	83%	80%	62%	96%	56%	98%

TAB. 1 – Tableau Comparatif entre $WAPC$ et d'autres classificateurs.

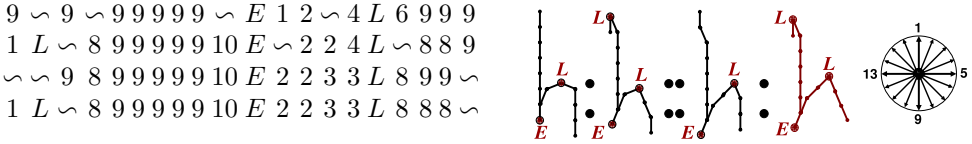

FIG. 1 – Resolution d'équations analogiques sur le code de Freeman augmenté pour la génération de caractères.

4.2 Reconnaissance de caractères manuscrits : génération de nouveaux exemples

Avec l'utilisation croissante du stylo dans l'interface homme machine, la reconnaissance des caractères manuscrits doit être de plus en plus précise et s'adapter rapidement aux nouveaux styles d'écritures (Mouchère *et al.* (2007)). La difficulté réside dans l'apprentissage avec très peu d'exemples. On peut augmenter efficacement l'ensemble d'apprentissage en générant automatiquement de nouveaux exemples. Pour ce problème, plusieurs stratégies de génération sont déjà employées. La première utilise la distortion d'image, la seconde est basée sur des transformations de la saisie on-line du scripteur (Mouchère & Anquetil (2006)). Nous proposons ici d'utiliser la proportion analogique pour la génération d'exemples.

Chaque caractère manuscrit est décrit par le code de Freeman à seize directions et enrichi par des points d'ancrages (représentés ici par des lettres majuscules). Les points d'ancrage sont par exemple des maxima ou des minima ou des variations angulaires. Ainsi, les caractères sont représentés par des séquences. En utilisant l'algorithme SOLVANA, on génère des nouvelles séquences synthétiques à partir de trois séquences originales (voir figure 1).

Afin de voir l'apport de la génération d'instance synthétique, nous avons appliqué les différentes méthodes de génération sur douze scripteurs qui ont saisi chacun 40 carac-

tères de chaque lettre de l'alphabet sur un PDA. Ayant 1040 caractères de chaque scripteur, nous avons pris 2,3, 4 et 6 caractères de chaque lettre de l'alphabet pour en générer 300 afin d'apprendre un classificateur (il s'agit d'une méthode RBFN). Enfin on calcule le taux de reconnaissance sur l'ensemble de test composé des caractères restant non utilisée en apprentissage (voir table 2). Pour mieux voir l'apport de la génération, notons que le taux de reconnaissance en utilisant 10 et 30 caractères pour l'apprentissage sans génération est respectivement 82.3% et 94.5%. En résumé, le nouveau scripteur peut écrire trois fois moins d'exemples pour avoir une qualité de reconnaissance égale.

Nb. de caractères utilisés	2	3	4	6
(1) : Distortions d'image	76.1 $\pm$ 3.3	82.5 $\pm$2.4	85.8 $\pm$2.0	89.4 $\pm$1.7
(2) : On-line et (1)	84.4 $\pm$2.6	88.0 $\pm$2.1	90.3 $\pm$1.7	92.3 $\pm$1.6
(3) : Analogie et (2)	84.9 $\pm$2.6	89.3 $\pm$2.1	91.3 $\pm$1.4	93.5 $\pm$1.1

TAB. 2 – Taux de reconnaissance moyen $\pm$écart type pour trois méthodes de génération

5 Conclusion

En conclusion, cet article a montré qu'une définition fondée de la proportion analogique et de la dissimilarité analogique, ainsi que leur mise en forme algorithmique, peuvent apporter de bons résultats dans deux domaines de l'apprentissage artificiel. Pour prendre un peu de recul, nous allons d'abord distinguer ce travail de ceux réalisés dans le cadre du CBR et de l'analogie en général. Ensuite, nous tenterons de dégager d'autres domaines pour lesquels cette approche pourrait se révéler intéressante.

En Intelligence Artificielle, le raisonnement par analogie est souvent, comme la citation de l'introduction l'a déjà évoqué, pratiquement confondu avec le CBR. Il s'agit de transférer de l'information d'un domaine bien connu vers un autre, en utilisant le fait que les deux domaines ont une structure commune (GENTNER *et al.* (2001)). Dans le cas particulier de la proportion analogique, les deux domaines sont confondus. Cette restriction a l'inconvénient de limiter les ambitions (en particulier sur la véracité cognitive du modèle), mais en revanche permet comme on l'a vu de donner un cadre formel et algorithmique fondé.

Pour en rester aux problèmes d'apprentissage de règles de classification, la question centrale est de savoir s'il est plausible que des objets en proportion analogique aient obligatoirement des classes également en proportion analogique. C'est en effet sur cette hypothèse que notre étude est fondée. Si l'on s'en tient, comme nous l'avons fait, à des proportions analogiques triviales dans les classes, le raisonnement est expérimentalement valable, comme le montre l'expérience du paragraphe 4.1. Il serait intéressant de le poursuivre en extrayant de l'ensemble d'apprentissage des proportions analogiques approchées, du genre "la classe 1 est à la classe 3 à peu près comme la classe 2 est à la classe 5", ou, dans le cas des caractères manuscrits, "a est à d à peu près comme o est à q". On réaliserait alors un système de transfert analogique entre l'univers des objets et celui des classes. Ceci a déjà été réalisé, par exemple dans YVON (1997), où les objets sont la forme orthographique de noms propres et les "classes" les formes phonétiques.

D'autre part, nous envisageons d'étendre nos études à des données structurées en arbre, par exemple pour l'apprentissage de la prosodie en synthèse de la parole.

Références

AAMODT A. & PLAZA E. (1994). Case-based reasoning : Foundational issues, methodological variations, and system approaches. *Artificial Intelligence Communications*, **7**(1), 39–59.

BAYOUDH S., MICLET L. & DELHAY A. (2007). Learning by analogy : a classification rule for binary and nominal data. In *Proc. of the Int. Joint Conf. on Artificial Intelligence*, volume 20, p. 678–683.

CORNUÉJOLS A. & MICLET L. (2002). *Apprentissage artificiel : concepts et algorithmes*. Eyrolles, Paris.

DASTANI M., INDURKHYA B. & SCHA R. (2003). Analogical projection in pattern perception. *Journal of Experimental and Theoretical Artificial Intelligence*, **15**(4).

GENTNER D., HOLYOAK K. J. & KOKINOV B. (2001). *The analogical mind : Perspectives from cognitive science*. Cambridge, MA : MIT Press.

HOFSTADTER D. & THE FLUID ANALOGIES RESEARCH GROUP (1994). *Fluid Concepts and Creative Analogies*. New York : Basic Books.

HOLYOAK K. J. (2005). Analogy. In *The Cambridge Handbook of Thinking and Reasoning*, p. 117–142 : Cambridge University Press.

LEPAGE Y. (2003). *De l'analogie rendant compte de la commutation en linguistique*. Grenoble. Habilitation à diriger les recherches.

LEPAGE Y. & ANDO S. (1996). Saussurian analogy : a theoretical account and its application. In *Proceedings of COLING-96*, p. 717–722, København.

MICLET L. & DELHAY A. (2005). *Analogical Dissimilarity : definition, algorithms and first experiments in machine learning*. Rapport interne 5694, INRIA.

MITCHELL M. (1993). *Analogy-Making as Perception*. Cambridge, MA : MIT Press.

MOUCHÈRE H., ANQUETIL E. & RAGOT N. (2007). Writer style adaptation in on-line handwriting recognizers by a fuzzy mechanism approach : The adapt method. *Int. Journal of Pattern Recognition and Artificial Intelligence*, **21**(1), 99–116.

MOUCHÈRE H. & ANQUETIL E. (2006). Synthèse de caractères manuscrits en-ligne pour la reconnaissance de lécriture. In *Actes du Colloque International Francophone sur l'Ecrit et le Document (CIFED'06)*, p. 187–192.

NEWMAN D., HETTICH S., BLAKE C. & MERZ C. (1998). UCI repository of machine learning databases.

SCHMID U., GUST H., KÜHNBERGER K.-U. & BURGHARDT J. (2003). An algebraic framework for solving proportional and predictive analogies. In R. Y. F. SCHMALHOFER & G. KATZ, Eds., *Proceedings of the European Conference on Cognitive Science (EuroCogSci 2003)*, p. 295–300, Osnabrück, Germany : Lawrence Erlbaum.

YVON F. (1997). Paradigmatic cascades : a linguistically sound model of pronunciation by analogy. In *Proceedings of the 35th annual meeting of the Association for Computational Linguistics (ACL)*, Madrid, Spain.

YVON F., STROPPA N., DELHAY A. & MICLET L. (2004). *Solving analogical equations on words*. Rapport interne ENST2004D005, École Nationale Supérieure des Télécommunications.