Méthode d'auto-fuzzyfication par analyse des typicalités sur des lots de données réduits

E. SCHMITT[1], V. BOMBARDIER[1], P. CHARPENTIER[1]

[1]Centre de Recherche en Automatique de Nancy, CNRS, UMR 7039, Faculté des Sciences

Boulevard des Aiguillettes, BP 239, 54506 Vandoeuvre-lès-Nancy Cedex

emmanuel.schmitt@cran.uhp-nancy.fr
vincent.bombardier@cran.uhp-nancy.fr
patrick.charpentier@cran.uhp-nancy.fr

Résumé – Cet article expose une méthode de fuzzyfication automatique pour un classificateur à base de règles linguistiques floues. Elle s'appuie sur l'analyse des scores de typicalité des attributs caractérisant les formes à classer. La méthode proposée est appliquée à la reconnaissance de couleur sur des avivés. L'utilisation d'un classificateur flou n'étant pas aisée pour des non experts, l'industrialisation d'une telle méthode nécessite une simplification des phases de réglages. En outre, le cadre applicatif spécifique de cette étude ne permet d'avoir à disposition qu'une quantité de données réduite pour réaliser la phase d'apprentissage. Les scores de typicalité des attributs présentent l'avantage de discriminer les plages de valeurs associées à chaque classe couleur de sortie. L'étude des corrélations de ces typicalités améliore la fuzzyfication des paramètres et les essais réalisés sur des lots de données « industrielles » montrent l'augmentation du taux de reconnaissance. Ces taux sont comparés à ceux obtenus à partir d'une fuzzyfication équirépartie. Par ailleurs, une diminution du nombre de règles floues générées dans le modèle est constatée. Les temps de traitements en généralisation sont ainsi réduits.

Abstract – This article exposes a method of automatic fuzzification, for a classifier based on fuzzy linguistic rules. It uses the analysis of the typicality scores of the attributes characterizing the patterns to be classified. The proposed method is applied to the color recognition on wooden boards. The use of a fuzzy classifier not being easy for non experts, the industrialization of such a method requires a simplification of the setting steps. Moreover, the specific industrial case of this study allows to have only tiny data sets to make the training step. The typicality scores of the attributes have the advantage to discriminate the variation spaces of the values associated to each output color class. The study of the typicality correlation improves the parameter fuzzification and the tests, which have been made on industrial data, show the improvement of the recognition rates. These rates are compared to those obtained with an equal distributed fuzzification. In addition, a reduction of the number of fuzzy rules generated in the model is noted. The processing times in generalization are thus reduced.

1. Introduction

Comme de très nombreuses industries manufacturières, les industries du bois sont placées sur un marché très concurrentiel. Elles expriment aujourd'hui le besoin de disposer de systèmes automatisés de vision de plus en plus adaptés au contrôle qualitatif du bois (détection de singularités, appariement colorimétrique).

Dans le domaine de classification d'avivés, plusieurs articles présentent l'utilisation des réseaux de neurones (RdN) [1][2]. De même, en classification, les Support Vecteur Machine (SVM) sont des techniques largement utilisées. Or, pour ces méthodes, le nombre d'échantillons doit être très important afin qu'elles puissent converger vers les classes de sortie souhaitées. Dans notre cadre applicatif, les industriels ne fournissent pas beaucoup d'échantillons qui sont pourtant indispensables à l'apprentissage des modèles de reconnaissance. Les classes de sorties sont alors incomplètes et hétérogènes à cause de la disparité du nombre de représentants de chaque classe.

Par ailleurs, notre problème est très lié au raisonnement humain. Les couleurs à classer sont intrinsèquement floues (transition progressive entre un *rouge foncé* et un *rouge moyen* sur une même planche impliquant une incertitude des caractéristiques extraites des images). Elles sont définies linguistiquement par les industriels ce qui implique une certaine subjectivité dans la définition des classes : la perception des couleurs à classer est alors considérée comme graduelle. En effet, les couleurs définies par les clients ne sont pas exprimées en termes de teinte, saturation, luminance (vocabulaire spécifique en traitement d'images) mais plutôt en termes d'aspect, de grain, de texture. Il existe donc un « fossé sémantique » entre les experts du domaine bois et ceux du domaine vision.

Les travaux présentés s'insèrent dans le cadre du développement d'un capteur flou couleur adapté à l'industrie du bois dont un des intérêts est de réduire ce fossé sémantique.

De même, dans un objectif d'amélioration de la production, le temps d'installation des systèmes automatisés (intégration mécanique du système, paramétrage du système, apprentissage des classes de sortie, …) doit être réduit au maximum. Les outils de réglages des systèmes développés n'étant pas toujours accessibles à des non experts, il est nécessaire d'en simplifier l'utilisation.

L'objet du présent article est de présenter une méthode originale de classification de plusieurs couleurs (*rouge*, *brun*, *blanc* par exemple) ainsi qu'une technique de réglage automatique des paramètres du classificateur. Deux aspects sont abordés dans cet article. Tout d'abord, les capacités de généralisation de la méthode sont démontrées suivant la quantité de données disponibles pour effectuer l'apprentissage. Puis une simplification de la phase de réglage du classificateur est proposée à partir de l'analyse des scores de typicalité des classes de sortie.

2. Capacité de généralisation

Pour les raisons évoquées précédemment, nous avons retenu un classificateur (FRC : Fuzzy Reasoning Classifier) basé sur un mécanisme de règles linguistiques floues [3] qui est adapté au contexte applicatif [4]. En effet, il possède une bonne capacité de généralisation et peut rendre compte d'une gradualité d'appartenance aux différentes classes de sortie. Nous avons choisi un mécanisme de règles conjonctives car elles sont générées à partir de données numériques extraites des images (caractérisation des couleurs du bois par moyennage des composantes colorimétriques) [5]. Le classificateur propose un module de génération automatique des règles basé sur un modèle de Larsen mettant en œuvre des règles floues du type « SI… ALORS… ». L'algorithme comporte trois parties : la fuzzyfication des attributs (vecteur caractéristique), la génération des règles et l'ajustement du modèle [6]. Le choix de la classe de sortie se fait alors en fonction de la règle à réponse maximale.

La faible quantité d'échantillons disponibles en apprentissage est une contrainte pour notre étude (classes de sortie caractérisées par 10 ou 20 échantillons en moyenne). En effet, l'analyse des taux de reconnaissance montre les difficultés de certaines méthodes classiques (RdN, SVM) à générer correctement des modèles à partir de peu d'échantillons. L'évolution du taux de reconnaissance en fonction du nombre de points en apprentissage est étudiée à partir d'un lot de données réelles (900 points représentant 6 classes). Pour comparer les capacités du FRC avec d'autres classificateurs, ces données ont été bruitées par un bruit blanc gaussien pour obtenir 5000 points par classe de sortie.

La figure 1 illustre la comparaison de l'efficacité de plusieurs classificateurs en fonction du nombre d'échantillons : classificateur bayésien, réseaux de neurones (RdN), k plus proches voisins (kppv), support vecteur machine (SVM) et classificateur flou (FRC). Quelque soit la méthode, les taux de reconnaissance tendent vers une asymptote horizontale : de 80% pour le classificateur bayésien, de 83% pour les k plus proches voisins (k=5), de 86% pour les réseaux de neurones et les SVM (non visible sur la courbe). Le FRC tend également vers une asymptote à 86%. Ces différentes méthodes convergent plus ou moins vite vers ces asymptotes. Typiquement, les réseaux de neurones et les SVM ont besoin de 1000 échantillons soit 40 fois plus de points en apprentissage pour atteindre des taux de reconnaissance semblables à ceux obtenus à partir du FRC. Ces résultats confirment la capacité du FRC à généraliser à partir de peu de données.

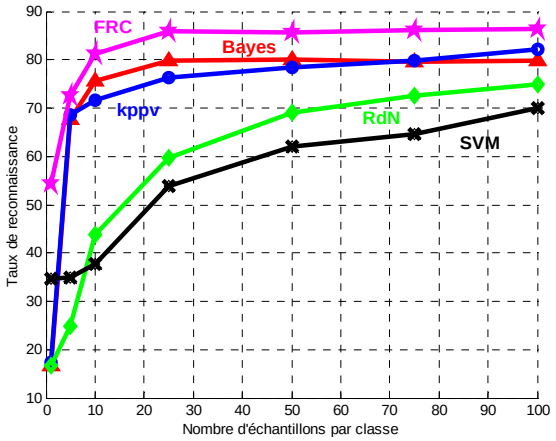

FIG. 1 – Evolution du taux de reconnaissance en fonction du nombre d'échantillons par classe en apprentissage

3. Auto-fuzzyfication des attributs

Dans un contexte industriel, il est important de simplifier les étapes de réglages des méthodes utilisées. Classiquement pour obtenir les meilleurs taux de reconnaissance, il est nécessaire de tester plusieurs configurations, ce qui représente souvent des heures de travail. Pour le FRC, cela concerne principalement l'étape de fuzzyfication des attributs. Cette partie des réglages représente la décomposition en plusieurs termes flous des composantes du vecteur caractéristique. Chaque terme ainsi créé correspond à un mot du langage naturel (l'intensité lumineuse peut être « claire », « moyenne » ou « foncée »).

Il existe trois manières de réaliser cette décomposition :

- soit par une fuzzyfication équirépartie (l'espace de variation du paramètre est découpé régulièrement en sous-espaces de même taille) ;
- soir par une fuzzyfication manuellement adaptée (l'univers de discours est découpé en fonction des connaissances *a priori* sur la discrimination des classes de sortie) ;
- soit par une fuzzyfication automatiquement adaptée (l'univers du discours du paramètre est découpé en fonction de la répartition des classes de sortie pour le paramètre considéré).

Les méthodes majoritairement utilisées pour effectuer une adaptation automatique de la fuzzyfication aux données d'apprentissage sont basées sur l'utilisation d'algorithmes génétiques [7] ou d'algorithmes de clustering [8][9]. L'inconvénient de ces méthodes réside dans le besoin de lots d'apprentissage conséquents.

Pour simplifier le réglage, les utilisateurs industriels préfèrent souvent une équipartition des termes. Pourtant, si le découpage ne correspond pas aux variations réelles des données, les termes créés sont inappropriés. La méthode d'auto-fuzzyfication présentée repose sur l'étude de la typicalité du vecteur caractéristique [10]. Le score de typicalité (1) des différents échantillons est calculé à partir des dissimilarités interclasses (2) et des ressemblances intraclasses (3).

$$T\left(V_a^u\right)=\frac{R\left(V_a^u\right).D\left(V_a^u\right)}{R\left(V_a^u\right).D\left(V_a^u\right)+\left(1-R\left(V_a^u\right)\right).\left(1-D\left(V_a^u\right)\right)} \quad (1)$$

$$D\left(V_a^u\right)=\frac{\sum_{i=1}^{m}\frac{1}{1-d\left(V_a^u,V_a^{e_i}\right)}}{m} \quad (2)$$

$$R\left(V_a^u\right)=\frac{\sum_{i=1}^{n}\frac{1}{d\left(V_a^u,V_a^{f_i}\right)}}{n} \quad (3)$$

où V_a^u est la valeur du paramètre a pour l'échantillon u

$V_a^{f_i}$ est la valeur du paramètre a pour l'échantillon f_i appartenant à la même classe que l'échantillon u

$V_a^{e_i}$ est la valeur du paramètre a pour l'échantillon e_i n'appartenant pas à la même classe que l'échantillon u

$d(\ )$ est la distance euclidienne

n et m sont respectivement les nombres d'échantillons appartenant ou n'appartenant pas à la classe de l'échantillon u.

La figure 2 représente les variations de ce paramètre pour une base de données industrielles composée de 6 classes de sorties (*Brun Foncé* - DB, *Brun* - B, *Brun Clair* - LB, *Rouge Foncé* - DR, *Rouge* - R et *Rouge Clair* - LR).

A partir des scores de typicalité entre les deux populations X et Y constituées chacune de n individus, le rapport $\rho_{corr/xcorr}$ entre le coefficient de corrélation et le coefficient de corrélation croisé des différentes classes a été calculé.

$$\rho_{corr/xcorr}\left(X,Y\right)=\frac{r\left(X,Y\right)}{\max_k\left(r_{cross}^k\left(X,Y\right)\right)} \quad (4)$$

où r est le coefficient de corrélation, r_{cross}^k est le coefficient de corrélation croisée à l'instant k.

Cette valeur correspond à un rapport entre le recouvrement et la ressemblance des courbes de typicalités pour chaque classe de sortie et pour chaque paramètre.

Les termes de fuzzyfication sont ensuite générés à partir de l'algorithme suivant :

1) Calculs des rapports $\rho_{corr/xcorr}$ pour toutes les classes de sortie

2) Ordonnancement croissant des classes de sorties en fonction de la valeur moyenne du paramètre considéré

i	1	2	3	4	5	6
Classe C_i	DR	DB	R	B	LR	LB

3) Création des termes de fuzzyfication :

POUR i = C_1 à C_K, **SI**

$$\rho_{corr/xcorr}\left(C_i,C_{i+1}\right)=\max\left[\rho_{corr/xcorr}\left(C_i,C_j\right)\,|\,j\in[1,K],j\neq i\right]$$

ALORS C_i et C_{i+1} sont représentés par le même terme **SINON** le terme en cours ne représente que la classe C_i, et un autre terme est créé pour la classe C_{i+1}.

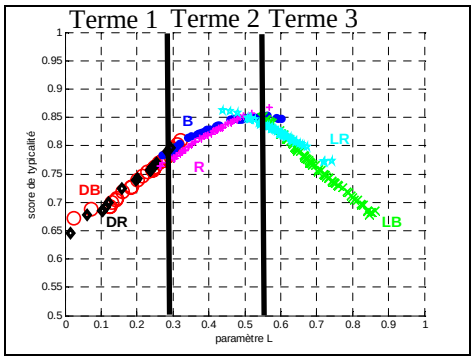

FIG. 2 : Score de typicalité pour le paramètre « Luminance » et fuzzyfication associée

En appliquant cet algorithme à l'exemple précédent (figure 2), le paramètre L a été fuzzyfié en 3 termes, les classes DB et DR constituant le premier terme, les classes R et B le deuxième terme, et les classes LR et LB le troisième terme.

Notre méthode se distingue des techniques classiques par son fonctionnement. En effet, elle se sert des classes de sorties pour effectuer la fuzzyfication automatique alors que les autres techniques se basent uniquement sur l'analyse des lots de données d'entrée (vecteurs caractéristiques). L'expertise des utilisateurs est donc exploitée dans le système.

4. Résultats

Les tests de validation de la méthode présentée ont été réalisés sur deux bases de données réelles constituées à chaque fois des mêmes 6 classes de sortie. Le tableau 1 illustre une comparaison de plusieurs techniques de fuzzyfication automatique : clustering [8][9], algorithme génétique (AG) [7] et analyse des typicalités [10].

TAB. 1 : Taux de reconnaissance des différentes méthodes de fuzzyfication automatique et nombre de règles générées

Base de données	Base de données 1*		Base de données 2**	
	Nb de règles	taux	Nb de règles	taux
Clustering	48	86.17%	60	85.56%
AG	36	85.37%	48	84.65%
Typicalités	**24**	**87.14%**	**45**	**86.10%**

(*943 points, 313 pour l'apprentissage et 630 pour la généralisation ; **1314 points, 209 pour l'apprentissage et 1105 pour la généralisation)

Les résultats obtenus à partir des différentes méthodes d'effectuer la fuzzyfication montrent deux améliorations : celle concernant le taux de reconnaissance et celle concernant le nombre de règles floues générées.

Tout d'abord, notre proposition améliore les taux de reconnaissance d'environ 1% par rapport aux autres techniques de fuzzyfication automatique. Ensuite, une diminution du nombre de règles apparaît avec l'utilisation de notre méthode de fuzzyfication. Ayant comme volonté la diminution de la complexité de la méthode de classification, il est important d'obtenir des bases de règles réduites pour qu'elles restent interprétables.

Outre la validation de notre technique de fuzzyfication automatique, le tableau 2 permet de valider le choix du FRC par rapport à des algorithmes de classification plus classiques.

TAB. 2 : Taux de reconnaissance des différentes fuzzyfication – Comparaison du FRC à de méthodes plus classiques

Base de données	Base de données 1	Base de données 2
Classificateur kppv (k=1)	81.37%	80.29%
Classificateur bayésien	79.12%	78.85%
Réseau de neurones	73.92%	71.25%
FRC – Fuzzyfication équirépartie (7 termes)	85.87% (343 règles)	84.9% (343 règles)
FRC – Fuzzyfication manuelle adaptée	84.28% (75 règles)	83.5% (75 règles)
FRC – Fuzzyfication automatique par typicalité	**87.14%** (24 règles)	**86.10%** (45 règles)

Les résultats du tableau 2 montrent la capacité du FRC à généraliser à partir d'un faible lot d'apprentissage puisque les taux de reconnaissance obtenus sont supérieurs à ceux obtenus par des méthodes classiques.

En outre, la fuzzyfication automatique que nous proposons permet, non seulement d'augmenter encore le taux d'environ 1.2%, mais surtout de diviser par un facteur 10 le nombre de règles du système d'inférences et ainsi de diminuer les temps de traitements. L'interprétabilité des modèles s'en voit donc simplifiée.

5. Conclusion

Le FRC basé sur un mécanisme de règles linguistiques floues présente un avantage certain concernant la capacité de généralisation des modèles. En effet, à partir de peu de points, il est possible d'obtenir des taux de reconnaissance satisfaisants. De plus, la technique d'auto-fuzzyfication proposée permet, d'une part, d'améliorer les taux de reconnaissance et, d'autre part, de réduire le nombre de règles générées. L'analyse des scores de typicalité des paramètres permet ainsi d'évaluer au mieux les recouvrements des différentes classes de sortie par rapport aux différents paramètres.

En se positionnant dans des problématiques plus complexes faisant intervenir un grand nombre de paramètres, il serait judicieux d'associer notre technique d'auto-fuzzyfication à une technique de sélection de paramètres [11] ce qui permettrait de garder toute l'interprétabilité des modèles linguistiques générés en diminuant la dimensionnalité du problème.

Références

[1] J. LAMPINEN, S. SMOLANDER, M. KORHONEN. Wood surface inspection system based on generic visual features. *International Conference on Artificial Neural Networks*, Paris, 1995.

[2] D. T. PHAM, S. SAGIROGLU. Training multilayered perceptrons for pattern recognition: a comparative study of four training algorithms. *International Journal of Machine Tools and Manufacture*, vol. 41, p. 419-430, 2001.

[3] L. A. ZADEH. Fuzzy sets. *Information and control*, vol. 8, p. 338-353, 1965.

[4] D. DUBOIS, H. PRADE. What are Fuzzy rules and how to use them?. *Fuzzy Sets and Systems*, vol. 84, p. 169-185, 1996.

[5] E. SCHMITT, V. BOMBARDIER, P. CHARPENTIER. Appariement de planches de bois par inférence floue. *20e colloque GRETSI sur le traitement du signal et des images*, p. 129, 2005.

[6] E. SCHMITT, C. MAZAUD, V. BOMBARDIER, P. LHOSTE. A Fuzzy Reasoning Classification Method for Pattern Recognition. *15th International Conference on Fuzzy Systems, FUZZIEEE'06*, Vancouver, Canada, p. 5998-6005 , 2006.

[7] O. CORDON, F. GOMIDE, F. HERRERA, F. HOFFMANN, L. MAGDALENA. Ten years of genetic fuzzy systems: current framework and new trends. *Fuzzy Sets and Systems*, vol. 141, p. 5-31, 2004.

[8] T. KEMPOWSKY, A. SUBIAS, J. AGUILAR-MARTI. Process situation assessment: From a fuzzy partition to a finite state machine. *Engineering Applications of Artificial Intelligence*, vol. 19, p. 461-477, 2006.

[9] F.A.T. DE CARVALHO. Fuzzy c-means clustering methods for symbolic interval data. *Pattern Recognition Letters*, vol. 28, p. 423-437, 2007.

[10] J. FOREST, M. RIFQI, B. BOUCHON-MEUNIER. Segmentation de classes pour l'amélioration de la construction de prototypes flous : visualisation et caractérisation de classes non homogènes. Rencontres Francophones sur la Logique Floue et ses Applications LFA, Toulouse, France, p. 29-36, 2006.

[11] C. MAZAUD, J. RENDEK, V. BOMBARDIER, L. WENDLING. A feature selection method based on Choquet integral and typicality analysis. *16th International Conference on Fuzzy Systems, FUZZIEEE'07*, Londres, Grande-Bretagne, 2007.