A propos des liens entre arbre de décision et treillis dichotomique

Karell Bertet, Stéphanie Guillas, Muriel Visani, Jean-Marc Ogier

A propos des liens entre arbre de décision et treillis dichotomique

Karell Bertet[1] – Stéphanie Guillas[1] – Muriel Visani[1] – Jean-Marc Ogier[1]

Laboratoire L3I - Univ La Rochelle - av M. Crépeau - 17042 La Rochelle

prenom.nom@univ-lr.fr

Résumé : *Dans ce papier, nous nous intéressons à la structure de treillis de Galois, treillis utilisé dans la méthode Navigala, méthode de reconnaissance de symboles basée sur un parcours (de type arbre de décision) dans le treillis, et plus généralement aux treillis dits* treillis dichotomiques*, définis à partir d'attributs binaires issus d'un traitement de discrétisation. Nous mettons en évidence les liens structurels unissant les arbres de décision et les treillis dichotomiques en montrant tout d'abord que tout arbre de décision est inclus dans le treillis, mais également que le treillis est en fait la fusion de tous les arbres de décision. Nous finissons par des expérimentations visant à comparer, pour de la reconnaissance de symboles, les performances des arbres de classification et des treillis construits avec la méthode Navigala.*

Mots-clés : « reconnaissance de symboles », « classification supervisée », « treillis de Galois », « arbre de décision »

1 Introduction

Depuis une vingtaine d'années, les treillis de Galois sont largement utilisés en classification supervisée [MEP 05], où ils donnent des résultats comparables à des méthodes connues de la littérature : ID3, C4.5, ... Leur structure à base de graphe nous a semblé pertinente pour une application dans le contexte de la reconnaissance de documents techniques. Ainsi, nous avons développé une méthode de classification dédiée aux symboles, appelée Navigala [GUI 07].

Le treillis est un graphe dont les noeuds, appelés concepts, sont des regroupements maximaux d'objets possédant le même sous-ensemble maximal d'attributs. La plupart des méthodes de classification à base de treillis [MEP 93, MEP 05, ZEN 04, SAM 04, OOS 88, VEN 97] utilisent ce graphe pour sélectionner des concepts. Les concepts ainsi sélectionnés serviront ensuite pour la classification qui est effectuée le plus souvent selon un principe de vote majoritaire.

La méthode Navigala se distingue : il s'agit non pas de sélectionner des concepts dans le treillis, mais d'utiliser la structure complète du graphe pour une navigation type arbre de décision. Elle propose ainsi plusieurs chemins vers une même classe, ce qui lui confère une meilleure robustesse que l'arbre de décision dans le cadre d'une classification de symboles détériorés. Dans cette méthode, les treillis de Galois possèdent deux particularités : la propriété de complémentarité par la borne supérieure et la propriété de co-atomisticité.

Ce papier s'intéresse aux treillis de Galois utilisés dans la méthode Navigala, méthode de reconnaissance de symboles basée sur un parcours au sein de ce graphe, et plus généralement aux treillis dits *treillis dichotomiques*, définis à partir d'attributs binaires issus d'un traitement de discrétisation.

Nous mettons en évidence les liens structurels entre treillis dichotomiques et arbres de décision définis à partir d'un même ensemble de données. Nous montrons que tout arbre de décision est inclus dans le treillis dichotomique, et que le treillis dichotomique est la fusion de tous les arbres.

En partie 2, nous donnons quelques définitions relatives au treillis de Galois et nous présentons de manière succinte la méthode de classification Navigala que nous avons développée. La partie 3 s'intéresse aux arbres de décision et aux treillis utilisés dans la méthode Navigala, et plus généralement aux treillis dichotomiques issus de données qui ont été discrétisées. Les liens structurels unissant les arbres de décision et les treillis dichotomiques sont étudiés et les preuves des différentes propositions sont établies. Nous finissons par quelques résultats expérimentaux dans la partie 4.

2 Navigala : méthode de classification de symboles par navigation dans un treillis de Galois

2.1 Définition d'un treillis de Galois

Un *treillis de Galois* ou *treillis des concepts* est défini à partir d'une relation binaire R entre un ensemble d'objets O et un ensemble d'attributs I appelée *table binaire*. On associe à un sous-ensemble d'objets $A \subseteq O$ l'ensemble $f(A)$ des attributs en relation avec tous les objets de A ; duallement on associe à un sous-ensemble d'attributs $B \subseteq I$, l'ensemble $g(B)$ de tous les objets en relation avec les attributs de B :

$$\begin{aligned} f(A) &= \{x \in I \mid pRx \; \forall \, p \in A\} \\ g(B) &= \{p \in O \mid pRx \; \forall \, x \in B\} \end{aligned}$$

Un *concept formel* est un sous-ensemble maximal objets-attributs en relation, défini formellement par un couple (A, B) avec $A \subseteq O$ et $B \subseteq I$, qui vérifie $f(A) = B$ et $g(B) = A$. On introduit alors la relation $\leq$ définie sur l'ensemble de tous les concepts formels par :
Pour deux concepts formels (A_1, B_1) et (A_2, B_2) :

$$(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow \left\| \begin{array}{l} A_1 \supseteq A_2 \\ \text{(équivalent à } B_1 \subseteq B_2) \end{array} \right.$$

La relation $\leq$ possède les propriétés d'une relation d'ordre, *i.e.* une relation transitive, antisymétrique et réflexive. Les propriétés d'une relation d'ordre permettent de considérer l'ensemble de tous les successeurs et de tous les

prédécesseurs d'un concept selon $\leq$. On peut également introduire les successeurs immédiats et prédécesseurs immédiats en considérant la relation de couverture de $\leq$ notée $\prec$.

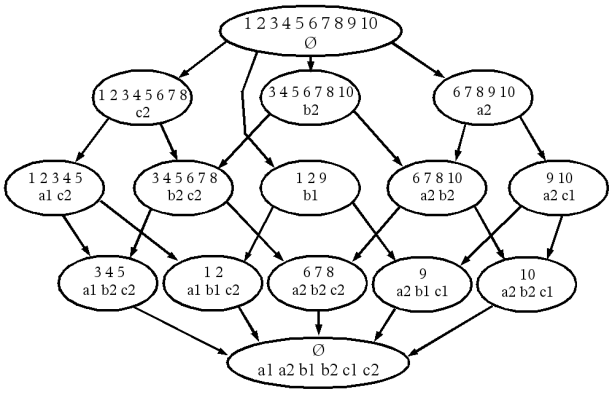
L'ensemble de tous les concepts formels équipé de la relation d'ordre $\leq$ est appelé *treillis de Galois* car il possède la propriété de treillis : pour tous concepts (A_1, B_1) et (A_2, B_2), il existe un unique plus grand successeur (resp. plus petit prédécesseur) appelé *borne inférieure* (resp. *borne supérieure*) noté $(A_1, B_1) \wedge (A_2, B_2)$ (resp. $(A_1, B_1) \vee (A_2, B_2)$) défini par :

$$(A_1, B_1) \wedge (A_2, B_2) = (g(B_1 \cap B_2), (B_1 \cap B_2)) \quad (1)$$

$$(A_1, B_1) \vee (A_2, B_2) = ((A_1 \cap A_2), f(A_1 \cap A_2)) \quad (2)$$

Cette propriété de treillis implique l'existence d'un unique plus petit élément $\perp = (O, f(O))$, et d'un unique plus grand élément $\top = (g(I), I)$. La figure 1 présente un exemple de treillis de Galois construit pour un ensemble de 10 objets décrits par 6 attributs ($a1$, $a2$, $b1$, $b2$,$c1$ et $c2$).

FIG. 1 – Treillis de Galois



Les deux fonctions f et g définies entre objets et attributs forment une *correspondance de Galois*. La composition $\varphi = f \circ g$, définie sur la famille des attributs, permet d'associer à un sous-ensemble d'attributs $X \subseteq I$ le plus petit concept contenant X : $(g(\varphi(X)), \varphi(X))$. Cette composition φ possède les propriétés d'un opérateur de fermeture : φ est idempotent (i.e. $\forall X \subseteq S$, $\varphi^2(X) = \varphi(X)$), extensif (i.e. $\forall X \subseteq S$, $X \subseteq \varphi(X)$) et isotone (i.e. $\forall X, X' \subseteq S$, $X \subseteq X' \Rightarrow \varphi(X) \subseteq \varphi(X')$). Pour plus d'informations sur le treillis de Galois et les opérateurs de fermeture, le lecteur peut se reporter aux références [BAR 70, GAN 99].

2.2 Description de la méthode Navigala

La méthode Navigala [GUI 07] a été conçue pour reconnaître des symboles issus de documents techniques. A partir des images de symboles, nous extrayons des signatures (vecteurs de caractéristiques). Nous avons implémenté trois signatures statistiques (Radon [TAB 03], Fourier-Mellin [DER 99] et Zernike [TEA 03]) que nous avons comparées dans [GUI 07]. Nous avons également développé une signature structurelle dédiée aux symboles [COU 07]. Cette signature est composée du nombre d'occurences de chemins dans un graphe topologique qui décrit les relations entre les segments préalablement extraits du symbole. Elle est invariante en translation, rotation, et changement d'échelle.

La méthode Navigala intègre les deux étapes classiques pour réaliser le processus de reconnaissance que sont l'apprentissage et la classification. Les données d'apprentissage sont tout d'abord discrétisées selon un critère de coupe supervisé, de manière à obtenir une table binaire. Les attributs de cette table discrétisée sont des intervalles de valeurs disjoints, et les objets sont les symboles. La discrétisation se termine lorsqu'il y a séparation des classes, c'est-à-dire lorsque chaque classe se distingue des autres par au moins un des attributs de la table. A partir de la table binaire discrétisée, la construction du treillis de Galois ne nécessite aucun paramètre et permet l'obtention d'une unique structure de graphe.

L'étape de classification de nouveaux symboles consiste ensuite à naviguer dans le treillis à la manière d'une navigation dans l'arbre de décision. Plus précisément, il s'agit, à partir du concept minimum, de progresser pas à pas d'un concept vers un successeur immédiat, en validant des intervalles de la table discrétisée, jusqu'à atteindre un concept final. Ce concept final permet d'attribuer une classe au nouveau symbole à classer, puisqu'il contient un ensemble d'objets appartenant tous à la même classe.

Notons que les intervalles utilisés pour valider le parcours au sein du treillis sont en réalité des ensembles flous, qui offrent plus de souplesse à la classification notamment dans le cas d'une base de test inconsistante. Enfin, il peut arriver que la base d'apprentissage ne permette pas d'obtenir une discrétisation avec séparation des classes ; dans ce cas, la classification sera plus imprécise étant donné que les concepts finaux pourront alors contenir des objets de classes différentes.

3 Treillis et arbre de décision

La classification mise en œuvre par la méthode Navigala est très proche de celle proposée par un arbre de décision. Dans cette section, nous revenons sur ces deux structures, et les données qu'elles manipulent.

3.1 Description d'un arbre de décision

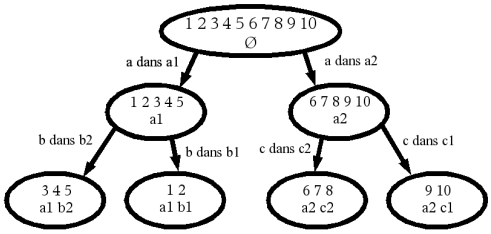
Depuis les années 1960-1970, l'arbre de décision construit à partir d'un ensemble de données a fait l'objet de nombreux travaux de recherche [RAK 97, RAK 05]. Les méthodes de génération de l'arbre de décision les plus connues sont ID3 [QUI 86], C4.5 [QUI 93] et CART [BRE 84].

Tout comme pour le treillis de Galois, les données d'entrée (discrètes, énumérées ou continues) sont représentées sous la forme d'une table contenant un ensemble d'objets décrits par un ensemble d'attributs.

Les noeuds de l'arbre sont construits depuis son sommet (la racine), vers sa base où les noeuds terminaux sont appelés feuilles. La construction de l'arbre nécessite un critère pour sélectionner, à chaque étape de division, une/des variable(s) de la table, un deuxième critère pour discrétiser les variables continues, et un critère d'arrêt des divisions généralement basé sur une mesure de pureté des feuilles. Le noeud-racine considère l'ensemble des objets de la table ; une variable de la table est alors sélectionnée pour partitionner ces objets

en deux sous-ensembles distincts formant ainsi deux noeuds fils. Le processus est réitéré sur chacun des noeuds créés, et ainsi de suite jusqu'à satisfaire le critère d'arrêt (voir Fig. 2).

FIG. 2 – Arbre de décision



Avec des données continues, comme dans le cas des signatures considérées ici, une discrétisation doit être réalisée :

- soit pendant la construction de l'arbre. Seules les variables sélectionnées pour la construction de l'arbre seront alors discrétisées.
- soit en prétraitement où il suffit de discrétiser les données jusqu'à la séparation des classes.

De nombreuses heuristiques sont envisageables pour réaliser la construction d'un arbre de décision. Il est par exemple très courant d'effectuer un élagage de l'arbre obtenu, de manière à éviter un sur-partitionnement des données. Le principe est de remonter à partir des feuilles de l'arbre en transformant certains noeuds de décision en feuilles selon un critère de pureté. Dans la comparaison structurelle décrite par la suite, les arbres de décision considérés ne sont pas élagués.

3.2 Treillis dichotomiques

Etant donné que la construction d'un arbre de décision induit une discrétisation des données, il est alors possible de considérer la table de données binaires issue de cette discrétisation, et par conséquent le treillis de Galois qui en résulte. On retrouvera ainsi dans cette table binaire, et dans le treillis, les mêmes attributs binaires que ceux proposés par l'arbre. Ainsi, si une variable V est proposée, avec deux fils, l'un pour *oui*, l'autre pour *non*, il s'agira de considérer les deux attributs binaires $V = oui$ et $V = non$. De façon plus générale, les attributs binaires issus des fils d'un noeud se retrouvent dans la table, et partitionnent l'ensemble des objets.

Dans la méthode Navigala, les variables sont des données continues qui sont discrétisées, et ce afin d'obtenir la séparation des classes. Les attributs binaires de la table sont les intervalles issus de cette discrétisation. Ainsi, un symbole sera décrit par une signature de longueur fixe avant discrétisation, puis par un ensemble d'intervalles binaires de même cardinalité après discrétisation. Un symbole sera associé à un seul des intervalles issus d'une même variable. Remarquons que les attributs binaires ainsi obtenus induisent une sélection automatique des variables discriminantes. En effet, un attribut partagé par tous les objets ne sera pas proposé dans l'arbre, et par conséquent ne sera pas pris en compte dans la table. C'est également le cas dans la méthode Navigala où une variable continue non discrétisée n'apparaît pas dans la table.

Lorsque tous les objets d'une table binaire sont associés à un même nombre d'attributs binaires, il en résulte que les concepts finaux (*i.e.* concepts correspondant à une classe) contiennent tous le même nombre d'attributs. Les concepts finaux du treillis ne peuvent donc pas être reliés entre eux, car deux concepts en relation selon $\leq$ ne peuvent être composés d'un même nombre d'attributs. Les concepts finaux ont alors pour unique successeur immédiat le concept $\top$. Cette propriété se retrouve en théorie des treillis sous le terme de *co-atomisticité*. C'est le cas dans la méthode Navigala.

Lorsque la discrétisation a lieu au cours de la construction de l'arbre, la table dépend alors des attributs proposés dans l'arbre, et deux arbres différents pourront engendrer deux ensembles d'attributs binaires différents. Ces deux ensembles d'attributs peuvent alors donner naissance à deux treillis différents. La discrétisation peut également être réalisée en prétraitement, comme dans la méthode Navigala. Même si plusieurs arbres de décision peuvent alors être construits à partir de cette table, un unique treillis lui sera associé.

Dans tous les cas, à tout attribut binaire x on peut associer un ensemble non vide $\overline{X}$ d'attributs binaires tel que les objets possédant l'attribut x, et ceux possédant les attributs de $\overline{X}$ sont tous différents. Les attributs binaires se déduisent de l'arbre : si x est une variable proposée par un noeud de l'arbre, alors $\overline{X}$ est l'ensemble de toutes les autres variables proposées par ce même noeud. Dans le cas de données continues discrétisées en prétraitement, x correspond à un intervalle, et $\overline{X}$ contiendra tous les autres intervalles pour cette même variable. De par cette propriété, les treillis issus d'un arbre correspondent à des treillis particuliers que nous appelons les *treillis dichotomiques*. Les treillis dichotomiques se caractérisent par le fait d'être $\vee$-complémentaires, c'est-à-dire que pour tout concept (A, B), il existe toujours un *concept complémentaire* (A', B') tel que :

$$(A, B) \vee (A', B') = \top = (\emptyset, I) \tag{3}$$

Proposition 1 *Tout treillis dichotomique (i.e. treillis issu d'un arbre) est $\vee$-complémentaire*

Preuve Soit (A, B) un concept d'un treillis dichotomique. Il s'agit de montrer l'existence d'un concept complémentaire à (A, B). Pour cela, considérons x un attribut binaire quelconque de B, et $\overline{x}$ un attribut complémentaire de x appartenant à l'ensemble $\overline{X}$. Il s'ensuit que les objets possédant x, et ceux possédant $\overline{x}$ sont différents, ce qui se formalise par $g(\{x\}) \cap g(\{\overline{x}\}) = \emptyset$. On considère alors le plus petit concept contenant $\overline{x}$ qui, par définition, sera le concept $(g(\varphi(\{\overline{x}\})), \varphi(\{\overline{x}\}))$ dont l'ensemble des attributs est $\varphi(\{\overline{x}\})$. On déduit de la définition des fonctions f et g que $g(\varphi(\{\overline{x}\})) = g(\{\overline{x}\})$, et que $A \subseteq g(\{x\})$. Etant donné que $g(\{x\}) \cap g(\{\overline{x}\}) = \emptyset$, on peut alors en déduire que $A \cap g(\varphi(\{\overline{x}\})) = \emptyset$. Par conséquent, $(A, B) \vee (g(\varphi(\{\overline{x}\})), \varphi(\{\overline{x}\})) = (\emptyset, I)$, et le concept $(g(\varphi(\{\overline{x}\})), \varphi(\{\overline{x}\}))$ est un concept complément de (A, B), ce qui prouve la $\vee$-complémentarité du treillis. □

3.3 Lien structurel entre treillis dichotomiques et arbres de décision

Un premier lien structurel entre l'arbre de décision et le treillis dichotomique réside dans le fait que ces deux struc-

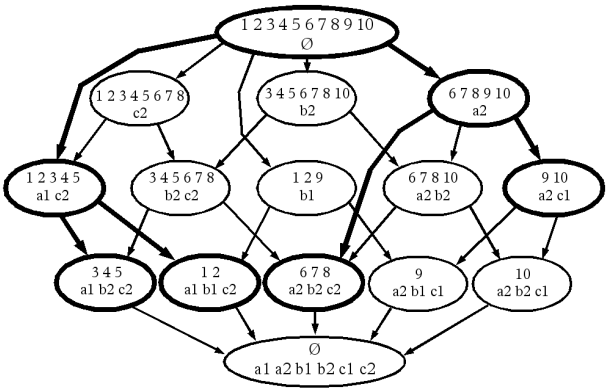
tures peuvent s'utiliser en classification supervisée, et être définies à partir d'une table d'attributs binaires.

On peut également noter une similarité entre les méthodes de classification supervisée basées sur une navigation dans le treillis et l'utilisation classique d'un arbre de décision. Cette similarité peut se formaliser par un lien structurel entre noeuds et concepts : en effet, il est possible d'associer à chaque noeud de l'arbre un unique concept dans le treillis. Considérons un noeud n de l'arbre, et l'ensemble des attributs binaires X_n proposés depuis la racine jusqu'à ce noeud. Sachant que ces attributs binaires se retrouvent dans la table à partir de laquelle le treillis est défini, on associe alors au noeud n le plus petit concept contenant les variables de X_n :

$$(g(\varphi(X_n)), \varphi(X_n)) \quad (4)$$

La figure 2 représente l'arbre de décision associé aux données de l'exemple. On peut noter que tous les nœuds de l'arbre de décision sont présents dans le treillis, et ce quelque soit le critère de construction de l'arbre. De plus, la structure de l'arbre (en gras) est incluse dans celle du treillis comme le montre la figure 3. Cette propriété se vérifie dans le cas général. Nous montrons ainsi que tout arbre de décision est inclus dans le treillis. Nous montrons également que le treillis est en fait la fusion de tous les arbres de décision.

FIG. 3 – Inclusion de l'arbre (en gras) dans le treillis



Proposition 2 *Tout arbre de décision est inclus dans le treillis dichotomique, lorsque ces deux structures sont construites à partir des mêmes attributs binaires.*

Preuve Considérons un arbre de décision et le treillis dichotomique issu des mêmes attributs binaires. A tout noeud n de l'arbre accessible par validation de l'ensemble d'attributs X_n on associe le concept $(g(\varphi(X_n)), \varphi(X_n))$ (cf. Eq. 4). Pour montrer que l'arbre est inclus dans le treillis, il s'agit alors de prouver les trois points suivants :

1. Deux noeuds différents d'un arbre de décision sont associés à des concepts différents :
   Si deux noeuds n_1 et n_2 sont associés au même concept, alors $\varphi(X_{n_1}) = \varphi(X_{n_2})$. Ceci signifie que ce sont les mêmes objets qui partagent les attributs de X_{n_1} et X_{n_2}, et donc que $n_1 = n_2$.

2. Si deux noeuds n_1 et n_2 sont ancêtres dans l'arbre alors leurs concepts associés sont en relation par $\leq$:
   Si n_1 est ancêtre de n_2, alors $X_{n_1} \subseteq X_{n_2}$. L'opérateur φ étant isotone, alors $\varphi(X_{n_1}) \subseteq \varphi(X_{n_2})$, et par conséquent les deux concepts $(g(\varphi(X_{n_1})), \varphi(X_{n_1}))$ et $(g(\varphi(X_{n_2})), \varphi(X_{n_2}))$ sont en relation selon $\leq$.

3. A l'inverse, si deux noeuds n_1 et n_2 ne sont pas ancêtres dans l'arbre de décision alors leurs concepts associés ne sont pas en relation dans le treillis :
   Si n_1 n'est pas ancêtre de n_2, considérons alors parmi les fils du plus petit ancêtre commun à n_1 et n_2, le fils n'_1 ancêtre de n_1 et le fils n'_2 ancêtre de n_2. n'_1 et n'_2 existent par construction de la table, et comme n'_1 et n'_2 sont frères, les attributs des concepts associés : $\varphi(X_{n'_1})$ et $\varphi(X_{n'_2})$, ne sont partagés par aucun objet. Ce qui se formalise par $g(\varphi(X_{n'_1})) \cap g(\varphi(X_{n'_2})) = \emptyset$. n'_1 étant ancêtre de n_1, on en déduit que $X_{n'_1} \subseteq X_{n_1}$, d'où $\varphi(X_{n'_1}) \subseteq \varphi(X_{n_1})$ par isotonie de l'opérateur φ, et à l'inverse $g(\varphi(X_{n'_1})) \supseteq g(\varphi(X_{n_1}))$ par définition de g. De même, $g(\varphi(X_{n'_2})) \supseteq g(\varphi(X_{n_2}))$ car n'_2 est ancêtre de n_2. D'où $g(\varphi(X_{n_2})) \cap g(\varphi(X_{n_1})) = \emptyset$, et par conséquent les concepts associés aux noeuds n_1 et n_2 ne sont pas en relation selon $\leq$.

□

Proposition 3 *Un treillis dichotomique est la fusion de tous les arbres de décision lorsque ces structures sont construites à partir des mêmes attributs binaires.*

Preuve Considérons un concept quelconque (A, B). Montrons que (A, B) est susceptible d'appartenir à un arbre de décision. On construit pour cela le sous-ensemble C de concepts du treillis contenant : le concept (A, B), un concept (A', B') complémentaire à (A, B), le concept minimal $\perp$, et tous les concepts finaux successeurs de (A, B) et de (A', B'). L'existence du concept complémentaire (A', B') se déduit de la propriété de $\vee$-complémentarité, induisant que ce sous-ensemble C équipé de la relation $\leq$ forme un arbre. On ajoute à l'ensemble C un nombre maximum de concepts du treillis dichotomique de sorte que $(C, \leq)$ reste un arbre. Nous obtenons ainsi un sous-arbre inclus dans le treillis dichotomique, contenant (A, B), et dont les feuilles, qui sont des concepts finaux, correspondent à des sous-ensembles d'objets qu'aucun attribut binaire ne peut séparer. Cet arbre peut donc être considéré comme un arbre de décision, ce qui termine cette preuve. □

4 Expérimentations

Les expérimentations ont été réalisées sur la base de symboles GREC 2003 [GRE]. Dans cette base, les déteriorations sur les symboles sont assimilables à celles obtenues par des appareils de reproduction (scanner, imprimante, ...). Chacune des 39 classes à disposition comporte 1 symbole modèle n'ayant subi aucune détérioration, et 90 symboles bruités. Les bases d'apprentissage et de test ne sont pas pré-définies.

TAB. 1 – Comparaison des taux de reconnaissance obtenus par l'arbre de décision et le treillis de Galois

	FM	Radon	Zernike
Arbre	59%	57%	43%
Treillis	71%	72%	59%

TAB. 2 – Complexité des structures

	FM	Radon	Zernike
Nb d'étapes de discrétisation	9	9	9
Nb d'intervalles discrétisés	18	17	18
Nb de noeuds de l'arbre	19	18	19
Nb de concepts du treillis	117	70	98

Nous avons tout d'abord comparé les taux de reconnaissance (Tab. 1) et les tailles des structures générées (Tab. 2) par l'arbre de décision (CART [BRE 84]) et le treillis de Galois, sur un extrait de 10 classes de la base GREC 2003. La base d'apprentissage était composée de 10 symboles (les symboles modèles) et la base de test de 900 symboles bruités. Les trois signatures statistiques : invariants de Fourier-Mellin (FM) [DER 99], R-signature (Radon) [TAB 03] et moments de Zernike [TEA 03], ont été étudiées pour ces tests. Pour chacune des signatures, le treillis de Galois est plus efficace que l'arbre de décision. La taille de l'arbre (polynomiale en la taille des données) est cependant plus condensée que celle du treillis (exponentielle dans le pire des cas, mais reste polynomiale en pratique dans les nombreuses expérimentations qui en ont été faites [MEP 05]). A la différence de l'arbre, le treillis propose plusieurs chemins de classification ou de reconnaissance, ce qui lui apporte une certaine robustesse lorsqu'il s'agit de reconnaître des données détériorées.

Pour remédier aux désavantages inhérents à la grande taille du treillis de Galois, nous avons réalisé une extension de l'algorithme de construction du treillis afin de ne générer à la demande que les concepts nécessaires à la navigation dans le graphe. Ainsi l'expérimentation suivante (Tab. 3), réalisée sur un extrait de 25 classes de la base GREC 2003, témoigne de l'apport d'une telle extension. La base d'apprentissage était composée de 25 symboles modèles et la base de test de 10 symboles bruités. Le nombre de concepts générés à la demande (282 concepts) est bien moins important que la construction du treillis entier (3185 concepts), et ce en effectuant le même parcours de reconnaissance dans le treillis. La génération à la demande garantit l'obtention de taux de reconnaissance identiques à ceux du treillis entier, tout en réduisant la taille de la structure à générer.

Nous avons également réalisé une comparaison du treillis de Galois avec d'autres classifieurs usuels : bayésien naïf, k-plus proches voisins et les SVM, sur un extrait de 10 classes de la base GREC 2003. Les taux de reconnaissance (voir Tab. 4) correspondent à la moyenne des taux obtenus en validation croisée à 5 blocs, 10 blocs et 26 blocs. Nous avons cependant inversé les ensembles d'apprentissage et de test par rapport à la validation croisée classique : ainsi, en validation 5 blocs, 1 bloc sert pour l'apprentissage (182 symboles) et les 4 autres pour le test (728 symboles), de manière à tester les limites de notre méthode pour des tailles d'apprentissage très faibles. L'approche Navigala obtient des taux relativement proches des autres classifieurs lorsque la taille de l'ensemble d'apprentissage est plus importante (5 blocs et 10 blocs), même si elle est moins performante que les autres approches. Avec moins de symboles en apprentissage (26 blocs), Navigala reste moins efficace que le k-ppv et les SVM, mais obtient un taux de reconnaissance plus élevé que le bayésien.

TAB. 3 – Génération du treillis entier/à la demande

	Apprent.	Classif.	Nb. concepts
Treillis entier	430,2 sec	2 sec	3185
Gén. demande	0,5 sec	9,8 sec	282

TAB. 4 – Comparaison des taux de reconnaissance obtenus par le treillis de Galois, le classifieur bayésien naïf, le k-ppv et les SVM à noyaux polynomiaux

Validation croisée	5 blocs Appr. 182 Reco. 728	10 blocs Appr. 91 Reco. 819	26 blocs Appr. 35 Reco. 875
Treillis	97,4%	92,2%	82,6%
Bayésien	99,4%	94,8%	62,8%
k-ppv (k=1)	100%	96,4%	90,5%
SVM (degré=1, échelle=2, offset=2,5)	100%	99,1%	93,6%

Dans cette dernière expérimentation, nous présentons les taux de classification obtenus par le treillis à partir de la signature de Radon et de la signature structurelle développée dans [COU 07]. Cette signature structurelle intègre une information complémentaire de celle proposée par les approches statistiques telle que la signature de Radon. Elle décrit l'organisation spatiale entre les primitives structurelles composant chaque symbole sous la forme d'un graphe topologique. Cette information spatiale est extraite du graphe topologique en calculant des chemins caractérisant des sous-structures spatiales (carré, losange, triangle, ...) incluses dans le symbole. Ces chemins sont disposés sous la forme d'un vecteur de valeurs, utilisé comme signature par la méthode Navigala. Dans cette expérimentation effectuée sur un extrait de 8 classes de la base GREC 2003, la base d'apprentissage était composée de 80 symboles (1 modèle + 9 symboles bruités par classe) et la base de test de 648 symboles bruités (81 symboles par classe). D'après les taux de reconnaissance obtenus (voir Tab. 5), la signature structurelle semble suffisamment fiable pour être utilisée conjointement avec des approches statistiques et améliorer les résultats actuels. Cette combinaison pourrait être mise en place par une reconnaissance hiérarchique itérative où la classification à une étape donnée serait raffinée à l'étape suivante.

TAB. 5 – Comparaison des taux de reconnaissance obtenus par le treillis de Galois selon la signature utilisée

	Radon	Structurelle
Treillis	98,9%	92%

Le treillis de Galois offre des perspectives intéressantes pour la reconnaissance d'objets détériorés. Dans ce cadre, sa structure apporte une plus grande robustesse à la classification que celle de l'arbre de décision.

Une des perspectives d'amélioration de Navigala se situe dans la phase de discrétisation. En effet, cette dernière est déterminante pour l'obtention d'un bon classifieur, puisque la construction du treillis est obtenue à partir de la table discrétisée et ne dépend d'aucun paramètre. Pour obtenir une meilleure discrétisation, nous souhaitons notamment nous inspirer du principe des SVM pour développer un nouveau critère de discrétisation.

5 Conclusion

Ce papier s'intéresse aux treillis de Galois utilisés dans la méthode de reconnaissance de symboles Navigala, basée sur un parcours dans un treilis de Galois, et plus généralement aux treillis dits *treillis dichotomiques*, définis à partir d'attributs binaires issus d'un traitement de discrétisation.

Parmi les différentes utilisations du treillis en classification supervisée, celle mise en place dans la méthode Navigala est basée sur une navigation dans le treillis, navigation similaire à l'utilisation classique d'un arbre de décision. Une telle navigation dans un treillis de Galois induit les mêmes avantages qu'un arbre de décision, à savoir la lisibilité du modèle et la capacité à sélectionner automatiquement les variables discriminantes parmi un très grand nombre de variables.

Notons cependant que la taille de l'arbre de décision est plus condensée que celle du treillis de Galois, mais que pour atténuer cet inconvénient le treillis peut être généré à la demande en cours de classification. A la différence de l'arbre, le treillis propose plusieurs chemins de classification ou de reconnaissance, ce qui lui apporte une certaine robustesse lorsqu'il s'agit de reconnaître des données détériorées.

Dans ce papier, nous décrivons les liens structurels qui unissent arbre de décision et treillis dichotomique en montrant que tout arbre de décision est inclus dans le treillis, mais aussi que le treillis est en fait la fusion de tous les arbres de décision. Nous donnons également quelques résultats expérimentaux mettant en valeur la similarité existant entre ces deux structures, ainsi que la plus grande robustesse du treillis par rapport à l'arbre.

6 Biliographie

Références

[BAR 70] BARBUT M., MONJARDET B., *Ordre et classification, Algèbre et combinatoire*, Paris, 1970, 2 tomes.

[BRE 84] BREIMAN L., FRIEDMAN J. H., OLSHEN R. A., STONE C. J., *Classification and regression trees*, Wadsworth Inc., Belmont, California, 1984.

[COU 07] COUSTATY M., GUILLAS S., VISANI M., BERTET K., OGIER J.-M., Flexible structural signature for symbol recognition using a concept lattice classifier, *Seventh IAPR International Workshop on Graphics Recognition (GREC'07)*, 2007.

[DER 99] DERRODE S., DAOUDI M., GHORBEL F., Invariant content-based image retrieval using a complete set of Fourier-Mellin descriptors, *Int. Conf. on Multimedia Computing and Systems (ICMCS'99)*, pp. 877-881, 1999.

[GAN 99] GANTER B., WILLE R., *Formal concept analysis, Mathematical foundations*, Springer Verlag, 1999.

[GRE] Base d'images GREC (Graphics RECognition), www.cvc.uab.es/grec2003/SymRecContest/ index.htm.

[GUI 07] GUILLAS S., Reconnaissance d'objets graphiques détériorés : approche fondée sur un treillis de Galois, PhD thesis, Université de La Rochelle, 2007.

[MEP 93] MEPHU NGUIFO E., Une nouvelle approche basée sur le treillis de Galois, pour l'apprentissage de concepts, *Mathématiques et Sciences Humaines*, vol. 124, pp. 19-38, 1993.

[MEP 05] MEPHU-NGUIFO E., NJIWOUA P., Treillis des concepts et classification supervisée, *Technique et Science Informatiques, RSTI*, vol. 24, n° 4, pp. 449-488, 2005.

[OOS 88] OOSTHUIZEN G., The use of a Lattice in Knowledge Processing, PhD thesis, University of Strathclyde, Glasgow, 1988.

[QUI 86] QUINLAN J., Induction of Decision Trees, *Machine Learning*, vol. 1, pp. 81-106, 1986.

[QUI 93] QUINLAN J. R., *C4.5 : Programs for Machine Learning*, Morgan Kaufman, Los Altos, California, 1993.

[RAK 97] RAKOTOMALALA R., Graphes d'induction, PhD thesis, Univ. C. Bernard, Lyon I, Décembre 1997.

[RAK 05] RAKOTOMALALA R., Arbres de décision, *Revue MODULAD*, vol. 33, pp. 163-187, 2005.

[SAM 04] SAMUELIDES M., ZENOU E., Learning-based visual localization using formal concept lattices, *2004 IEEE Workshop on Machine Learning for Signal Processing*, pp. 43-52, 2004.

[TAB 03] TABBONE S., WENDLING L., Adaptation de la transformée de Radon pour la recherche d'objets à niveaux de gris et de couleurs, *Technique et Science Informatiques*, vol. 22, n° 9, pp. 1139-1166, 2003.

[TEA 03] TEAGUE M., Image analysis via the general theory of moments, *Journal of Optical Society of America (JOSA)*, vol. 70, pp. 920-930, 2003.

[VEN 97] VENTER F., OOSTHUIZEN G., ROOS J., Knowledge discovery in databases using lattices, *Expert Systems With Applications*, vol. 13, n° 4, pp. 259-264, 1997.

[ZEN 04] ZENOU E., SAMUELIDES M., Utilisation des treillis de Galois pour la caractérisation d'ensembles d'images, *RFIA'04*, vol. 1, pp. 395-404, 2004.