Classificateurs aléatoires Topologiques à base de graphes de voisinage

Fabien Rico*, Djamel A. Zighed*

*Laboratoire Eric, Univ. de Lyon, 5, av. P. Mendès France, 69676 Bron Cedex, France
fabien.rico@univ-lyon1.fr & zighed@univ-lyon2.fr
http://eric.univ-lyon2.fr

Résumé. En apprentissage supervisé, les Méthodes Ensemble (ME) ont montré leurs qualités. L'une des méthodes de référence dans ce domaine est les Forêts Aléatoires (FA). Cette dernière repose sur des partitionnements de l'espace de représentation selon des frontières parallèles aux axes ou obliques. Les conséquences de cette façon de partitionner l'espace de représentation peuvent affecter la qualité de chaque prédicteur. Il nous a semblé que cette approche pouvait être améliorée si on se libérait de cette contrainte de manière à mieux coller à la structure topologique de l'ensemble d'apprentissage. Dans cet article, nous proposons une nouvelle ME basée sur des graphes de voisinage dont les performances, sur nos premières expérimentations, sont aussi bonnes que celles des FA.

1 Introduction

En data mining, plus particulièrement en apprentissage supervisé, on a de plus en plus recours à la combinaison d'un ensemble de prédicteurs. Le principe consiste à générer un ensemble de classifieurs selon une ou plusieurs méthodes d'apprentissage données et d'agréger ensuite leur prédiction par un vote.

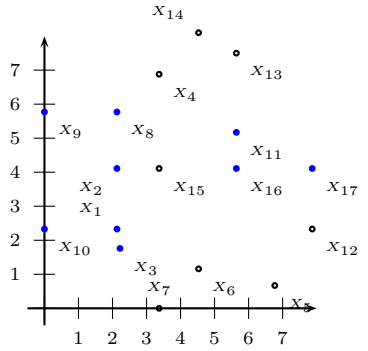
Le principe commun à de nombreuses ME est que, l'espace des observations est découpé en régions soit de manière quasi aléatoire comme le proposent [Kleinberg (2000, 1996); Ho (1998); Ho et Kleinberg (1996); Fradkin et Madigan (2003)] soit selon des règles générales fixées par un algorithme comme les arbres de décision comme le fait [Breiman (2001)]. Chaque région constitue ensuite un classifieur local.

Il nous semble que la manière dont sont construites ces régions doit influer sur chaque classifieur individuel et par conséquent peut affecter le classifieur agrégé. Nous proposons un nouveau moyen de construire une telle méthode basé sur l'utilisation de graphes de voisinage. Cette méthode présente d'ores et déjà des performances comparables aux Forêts Aléatoires (FA) considérées souvent comme très performantes.

La section suivante va donner les principales définitions et formaliser les méthodes ensemble en prenant comme exemple les forêts aléatoires. Dans la section 3, nous introduisons l'utilisation des graphes de voisinage. La section 4 montrera nos premiers résultats comparés à ceux des forêts aléatoires.

2 Concepts de base

On considère un échantillon d'apprentissage E_a composé de n individus $(\omega_i)_{1...n}$, décrits par p variables prédictives $X^j, j = 1, \ldots, p$ et une classe d'appartenance Y. La description d'un individu est $X(\omega_i) \in \Re$ et la classe $Y(\omega_i) \in \{c_1, \ldots, c_K\}$. On se place ainsi, dans le cadre d'un problème d'apprentissage supervisé à K classes. Pour simplifier les notations, on note $X_i = X(\omega_i)$ et $Y_i = Y(\omega_i)$. Considérons l'exemple suivant dans $\mathbb{R}^2$ avec une variable de classe binaire, $c_1 = 1$ ou $c_2 = 2$ sur la figure 1. Nous utiliserons cet exemple jouet pour illustrer différentes définitions.



E_a	X^1	X^2	Y	E_a	X^1	X^2	Y
X_1	2,13	2,33	2	X_{10}	0	2,33	2
X_2	2,13	4,11	2	X_{11}	5,64	5,17	2
X_3	2,22	1,76	2	X_{12}	7,87	2,33	1
X_4	3,37	6,88	1	X_{13}	5,64	7,5	1
X_5	6,77	0,67	1	X_{14}	4,53	8,1	1
X_6	4,53	1,16	1	X_{15}	3,37	4,11	1
X_7	3,37	0	1	X_{16}	5,64	4,11	2
X_8	2,13	5,77	2	X_{17}	7,87	4,11	2
X_9	0	5,77	2				

FIG. 1 – *Ensemble de points dans R^2 avec deux classes*

2.1 Classifieurs simples

Dans tout problème d'apprentissage supervisé multiclasses, pour un individu ω représenté par $X = X(\omega)$ on cherche à prédire correctement la classe $Y = Y(\omega)$. Pour y parvenir, on cherche à calculer la distribution de probabilité $P(Y/X) = (p(Y = c_k/X); k = 1, \ldots, K)$.

Définition 1 (Classifieur) *Dans un espace de représentation $\Re$ de dimension p par exemple $\mathbb{R}^p$, un* classifieur *est une application de $\Re$ dans le simplexe d'ordre K (où K est le nombre de classes), $\phi : X \longmapsto (p_1, p_2, \ldots, p_K)$*

Nous nous intéresserons plus particulièrement à un ensemble de classifieurs que nous appellerons *classifieurs topologiques* et qui sont des classifieurs utilisant les informations de proximité entre les individus.

2.2 Classifieur Topologique

Un Classifieur Topologique (CT) est un classifieur construit à partir de l'ensemble d'apprentissage $E_a \subset \Re$ qui repose sur la notion de proximité. C'est à dire que la première opération pour classifier un point X sera de lui associer un *voisinage* dans l'ensemble d'apprentissage. Ce dernier permettra de calculer le résultat $\phi(X)$ grâce à un *classifieur individuel local*. Ce résultat ne tiendra compte que de X et de son voisinage dans E_a.

Différents exemples de classifieurs topologiques peuvent être trouvés dans la littérature : les $k-$plus proches voisins, les $\varepsilon-$voisins, les arbres de décision, les graphes de voisinage (GABRIEL, voisins relatifs, arbre de longueur minimale, polyèdres de DELAUNAY,...).

1. ***La base de voisinage*** $\mathcal{P}$ ***:*** c'est un ensemble de parties de E_a qui forment un recouvrement de l'ensemble d'apprentissage. Cette base est formée de tous les voisinages possibles. Tout individu à classifier sera ainsi rattaché à l'un des éléments de cette base.
2. ***La fonction de voisinage*** $\mathcal{V}$ ***:*** permet de rattacher un individu à un élément de la base de voisinages. Cette fonction permet d'associer à tout point X un sous-ensemble de E_a qui sont ses voisins. Seuls ces points voisins interviendront dans la détermination de la classe de Y.
3. ***Un classifieur local*** qui permet d'estimer localement la distribution de probabilité des classes : $C : \Re \times \mathcal{P} \longrightarrow S_K$

D'où la définition d'un classifieur topologique ϕ formé par le triplet $(\mathcal{P}, \mathcal{V}, C)$:

$$\phi(X) = C_{\mathcal{V}(X)}(X)$$

2.3 Méthode Ensemble

Le principe des Méthodes Ensemble est de générer M classifieurs et de les agréger en un seul. Pour cela, on effectue M itérations, toutes identiques. A l'itération m :

1. On génère un nouvel ensemble d'apprentissage E_a^m dans un espace de représentation $\Re^m$. Pour cela, il est par exemple possible d'échantillonner les données de départ, avec ou sans remise et/ou en projetant les données.
2. On génère un nouveau classifieur $\phi^m = (\mathcal{P}^m, \mathcal{V}^m, C^m)$ éventuellement en rajoutant un aléa (par exemple par la sélection des variables dans les random forest)
3. On agrège les résultats obtenus par exemple par la majorité simple (random forest) ou la majorité pondérée.

2.4 Exemple des forêt aléatoire

– ***Créations des ensembles d'apprentissages :*** échantillonage bootstrap de l"ensemble de départ
– ***Base de voisinage*** $\mathcal{P}$ ***:*** lorsqu'on construit l'arbre, on engendre en fait une partition de l'ensemble d'apprentissage qui constitue la base de voisinage.
– ***Fonction de voisinage*** $\mathcal{V}$ ***:*** quand un nouvel individu X se présente, il est facile de localiser son voisinage en parcourant l'arbre de la racine au noeud terminal. Cela induit également une partition de l'espace de représentation comme on le voit dans la figure 2 où les différentes régions sont délimitées par les droites perpendiculaires aux axes.
– ***Classifieur local :*** il consiste alors à prendre la proportion de chaque classe dans le voisinage.
– ***agrégation :*** un vote à la majorité simple.

La figure 2 propose un arbre de classification qui engendre une partition de $\mathbb{R}^2$ en 5 régions.

3 Utilisation des graphes de voisinage

À aucun moment, dans les forêts aléatoires, nous n'avons exploité la notion de région au sens géométrique du terme. En fait, nous aurions pu partitionner l'espace selon des formes

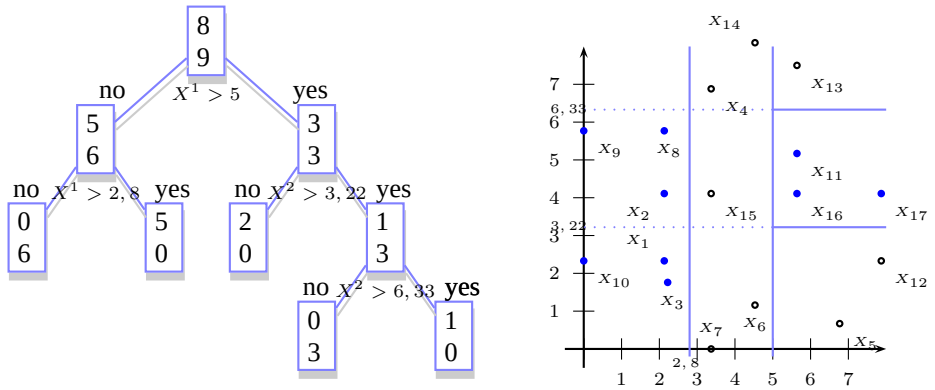


FIG. 2 – *Partitionnement de l'espace de représentation par un arbre*

quelconques. La seule difficulté, qu'il faut garder à l'esprit, est qu'il faut être en mesure de localiser la région de tout individu dont on veut prédire la classe. Pour cela nous devons construire une base de voisinages $\mathcal{P}$ et une fonction de voisinage $\mathcal{V}$.

3.1 Partitionnement par graphes de voisinage

Dans cet article, nous nous sommes limités aux structures de voisinages géométriques issues des graphes de GABRIEL ([Park et al. (2006)]) Le graphiques 3 montre, sur le jeu de données introduit plus haut (cf §2), le graphe de GABRIEL.

- ***Créations des ensembles d'apprentissages :*** nous échantillonnons les données de départ avec remise et les variables sans remise.
- ***Base de voisinage*** $\mathcal{P}$ ***:*** à partir du graphe de GABRIEL, il suffit pour cela de couper les arrêtes qui relient des sommets n'appartenant pas à la même classe. Les composantes connexes de ce graphe réduit forment la base de voisinages.
- ***Fonction de voisinage*** $\mathcal{V}$ ***:*** il faut définir une notion d'attractivité d'une partie de E_a sur un individu. Pour cela on peut utiliser la propriété du graphe de GABRIEL. L'attractivité de P sur x serait :

$$\begin{aligned} a(P/x) &= \#\left(\{y \in P / R_{GG}(x,y)\}\right) \\ \text{où } R_{GG}(x,y) = 1 &\Leftrightarrow d(x,y) \leq \sqrt{d^2(x,z) + d^2(z,y)}; \forall z \in E_a \end{aligned}$$

  Un individu anonyme est raccroché à la composante dont l'attractivité est la plus grande.
- ***agrégation :*** le voisinage de chaque point étant pure, il attribut au points sa classe, pour l'agrégation de plusieurs classifieurs, nous avons utiliser comme pondération la la taille de la composante de rattachement (c'est à dire le cardinal du voisinage).

3.2 Intérêts des graphes de voisinage

Avec les Forêts Aléatoires, le partitionnement de l'espace est contraint car en utilisant les coupures parallèles aux axes ou obliques on obtient forcement des polyèdres convexes. Nous

avons observé que ces contraintes peuvent avoir un effet négatif sur la capacité de généralisation. Sur des formes non linéairement séparables telles que des spirales imbriquées, les performances sont bien meilleures pour les graphes que pour les FA (6% d'erreur au lieu de 30%).

L'autre avantage de notre approche est que la construction d'un graphes de voisinage, au contraire de celle d'un arbre, n'est pas liée à un espace de représentation. Elle peut se faire à partir d'une simple matrice de distance ou de similarité. Cela permet d'utiliser cette méthode d'apprentissage sur des jeux de données (par exemple issus des sciences sociales) pour lesquels il est parfois difficile de trouver un bon espace de représentation.

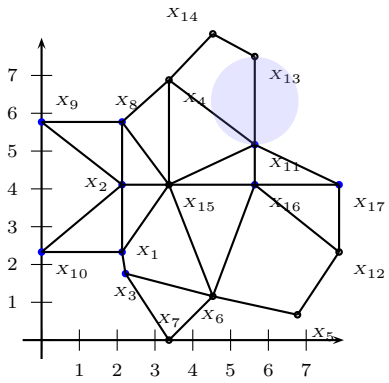


FIG. 3 – *Voisinage Graphe de* GABRIEL

4 Évaluation

Pour évaluer l'intérêt des graphes de voisinage, nous allons utiliser des jeux de données du site de l'UCI. Dans ces premières expérimentations, nous nous sommes limités aux jeux de données quantitatifs avec un petit nombre de classes. Nous avons conduit les mêmes tests en utilisant les forêts aléatoires (selon la version implémentée pour R dans la librairie `randomForest` [Liaw et Wiener (2002)]) les paramètres donnant le meilleur résultat dans ce cas était généralement les paramètres par défaut ($mtry = \sqrt{d}$ où d est la dimension et $ntree = 500$). Chaque test à été répété 100 fois de manière à obtenir un estimation du taux d'erreur et un intervalle de confiance pour les 2 méthodes. Ces intervalles de confiances ont été calculés avec un risque d'erreur de première espèce $\alpha = 0.99$ en utilisant un test de student. Les résultats sont consignés dans le tableau 1.

Jeu de données	N appr.	N test	p var.	K classes	CAT/GABRIEL		Forêt aléatoire	
					% Err. Moy.	Intervalle Confiance	% Err. Moy.	Intervalle Confiance
Waveform	2500	2500	21	3	14.70	[14.26, 15.14]	14.68	[14.22, 15.14]
Twonorm	1064	1064	20	2	2.70	[2.61, 2.80]	3.39	[3.26, 3.52]
Trinorm	1064	1064	20	2	13.46	[13.23, 13.69]	14.44	[14.23, 14.65]
Ringnorm	1064	1064	20	2	4.36	[4.07, 4.65]	5.25	[4.88, 5.61]
Sonar	150	58	60	2	18.12	[16.71, 19.54]	16.34	[14.75, 17.94]
Diabetes(Pima)	384	384	8	2	32.19	[31.35, 33.01]	23.91	[23.00, 24.83]
Ionosphere	280	70	34	2	7.60	[6.82, 8.39]	6.59	[5.83, 7.35]
Letters(RvsB)	760	760	16	2	2.38	[2.19, 2.56]	2.59	[2.41, 2.76]
Musk2	2000	2000	166	2	7.70	[7.32, 8.08]	5.06	[4.65, 5.47]

TAB. 1 – *Comparaison des CAT et des FA sur plusieurs jeux de l'UCI*

En comparant les résultats des deux méthodes on s'aperçoit que les performances des CAT basées sur les graphes de GABRIEL sont en moyenne équivalentes à ceux des forêts. Sur les 9 jeux de données, CAT/GABRIEL est meilleur que les Forêts Aléatoires 4 fois et est équivalent à 1 reprises.

5 Conclusions et travaux futurs

Nous venons de proposer un cadre qui permet d'unifier les approches Ensemble utilisant la notion de voisinage dans un cadre plus large, celui des classifieurs aléatoires topologiques. Nous avons montré comment cette approche peut être instanciée soit par les Forêts Aléatoires, soit de manière analogue par les graphes de voisinage. Dans les méthodes ensemble le recourt à une meilleure exploitation de la structure topologique de l'ensemble d'apprentissage ouvre des pistes qui nous paraissent prometteuses. On peut en effet les utiliser dans des situations où l'espace de représentation n'est pas explicite mais où seule la matrice de proximité est connue. Nous faisons ainsi une connexion avec les méthodes à base de noyaux.

Références

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Fradkin, D. et D. Madigan (2003). Experiments with random projections for machine learning. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 522. ACM.

Ho, T. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*(8), 832–844.

Ho, T. et E. Kleinberg (1996). Building projectable classifiers of arbitrary complexity. In *International Conference on Pattern Recognition*, Volume 13, pp. 880–885.

Kleinberg, E. (1996). An overtraining-resistant stochastic modeling method for pattern recognition. *The annals of statistics 24*(6), 2319–2349.

Kleinberg, E. (2000). On the algorithmic implementation of stochastic discrimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*(5), 473–490.

Liaw, A. et M. Wiener (2002). Classification and regression by randomforest. *R news 2*(3), 18–22.

Park, J., H. Shin, et B. Choi (2006). Elliptic Gabriel graph for finding neighbors in a point set and its application to normal vector estimation. *Computer-Aided Design 38*(6), 619–626.

Summary

In supervised machine learning, Ensemble Methods (EM) are known to perform better than single classifier methods. One of the main reference method is the Random Forest. This method shatters the feature space according to hyperplans wich are either parallel to the axis or oblic. The consequencies of this partitionning might affect the performances of the classifier. We believe that this approach could be improved by better exploiting the topology of the dataset. For that purpose, we use neighbourhood graphs.