Utilisation des graphes de proximité dans le cadre de l'apprentissage basé sur les voisins

Sylvain Ferrandiz*, Marc Boullé**
*2 avenue Pierre Marzin, 22307 Lannion
sylvain.ferrandiz@rd.francetelecom.com,
**2 avenue Pierre Marzin, 22307 Lannion
marc.boulle@rd.francetelecom.com,

Résumé. La classification suivant les plus proches voisins est une règle simple et attractive, basée sur une définition paramétrique du voisinage. Les graphes de proximité, quant à eux, induisent des notions plus souples de voisinage. Il s'agit ici d'effectuer la substitution.

Les variantes obtenues, peu testées dans la bibliographie, ont été soumises à une expérimentation intensive, sur bases de données de l'UCI et de France Télécom. On a ainsi considéré divers types de prétraitement des données et plusieurs catégories de graphes. De plus, on a caractérisé les effets du "piège de la dimension" sur le comportement théorique de tous les graphes présentés, une quantification empirique du phénomène ayant été réalisée.

Il ressort de notre étude que l'utilisation du voisinage de Gabriel provoque une amélioration en moyenne et que le prétraitement basé sur la statistique de rang est le plus adéquate. Quoiqu'il arrive, des précautions doivent être prises en grande dimension.

1 Introduction

La classification constitue un problème d'apprentissage classique qui se présente comme suit. On cherche à prédire la valeur d'un attribut cible (ou variable endogène), prédiction basée sur un ensemble S de n instances structurées en vecteurs d'un produit cartésien de d attributs descriptifs (ou variables exogènes) et un ensemble E d'étiquettes correspondantes. On suppose ici les attributs descriptifs tous à valeurs réelles.

Une méthode de classification, proposée dans [Fix et Hodges, 1951], est la règle k-NN. Elle consiste, pour une instance x à étiqueter, à effectuer un vote à la majorité sur les k plus proches voisins (au sens euclidien) de x dans S. Un résultat de [Cover et Hart, 1967], étendu à toutes distributions par Stone et Devroye, affirme que le plus proche voisin de x contient plus de la moitié de l'information sur x. Autrement dit, si L^* désigne l'erreur bayésienne (optimale) et $L_{1-\mathrm{NN}}$ l'erreur de la règle 1-NN, on a $L^* \leq L_{1-\mathrm{NN}} \leq 2L^*$.

De plus, la règle k-NN, dans le cas à deux modalités cibles, est asymptotiquement optimale [Stone, 1977]. Ainsi, si n devient infini et si la suite (k_n) du nombre de voisins pris en compte tend vers l'infini suffisament lentement par rapport à n (i.e $k_n/n \to 0$), l'erreur $L_{\mathrm{k_n}-\mathrm{NN}}$ converge en probabilité vers L^*.

La classification repose donc sur le voisinage de l'instance à étiqueter. Dans la règle k-NN, celui-ci est défini de manière globale et paramétrique. L'introduction des graphes