Comparaison de modèles linéaires pour déterminer la distance optimale pour un apprentissage par plus proches voisins

Yuzhen Wang[1,2] Pierre Lemaire[1] Iragaël Joly[2]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, G-SCOP, 38000 Grenoble, France
[2] Univ. Grenoble Alpes, INRA, Grenoble INP, GAEL, 38000 Grenoble, France
(nom.prenom@grenoble-inp.fr)

Mots-clés : *programmation linéaire, k-NN, distance euclidienne pondérée*

1 Présentation du problème

L'algorithme des plus proches voisins (*k-nearest neighbors* ou k-NN) est un algorithme classique d'apprentissage supervisé : une observation non-classifiée est classée selon la classe majoritaire de ses k plus proches voisins [4].

Le plus souvent dans ses applications, la distance est simplement la distance euclidienne standard sur l'ensemble des attributs (ou variables) des observations. Il est toutefois possible d'améliorer les performances de k-NN en utilisant d'autres distances, en particulier des distances euclidiennes pondérées. Pour cela, Hocke et Martinetz [2] ont proposé 2 modèles linéaires pour déterminer les poids des attributs (soit $x_{i,m}$ le jeu de donnée et $a_{i,j,m} = (x_{i,m} - x_{j,m})^2$) :

$$\begin{array}{lllll} \min & d_{intra} & & & \\ s.t. & \sum_{m=1}^{M} w_m\, a_{i,j,m} & \leq & d_{intra} & (y_i = y_j) \\ & \sum_{m=1}^{M} w_m\, a_{i,j,m} & \geq & 1 & (y_i \neq y_j) \end{array} \qquad \begin{array}{lllll} \min & d_{intra} + C\sum_{i=1}^{N} \xi_i & & & \\ s.t. & \sum_{m=1}^{M} w_m\, a_{i,j,m} & \leq & d_{intra} + \xi_i & (y_i = y_j) \\ & \sum_{m=1}^{M} w_m\, a_{i,j,m} & \geq & 1 & (y_i \neq y_j) \end{array}$$

Le modèle de gauche est la version «basique» qui détermine les poids w_m de manière à minimiser la distance d_{intra} entre deux observations d'une même classe, sous contrainte que des observations de classes différentes soient au moins à distance 1 (cette valeur, arbitraire, normalise les poids). Le modèle de droite est une version «souple» avec des tolérances (variables $\xi_i \geq 0$) qui autorise des erreurs pénalisées via un paramètre C dont la valeur est déterminée par l'essai successif de 11 valeurs candidates ($C \in \{2^{-q}, q = 0..10\}$) selon Hocke et Martinetz [2].

À partir de ces deux modèles, nous avons proposé différentes variantes en modifiant une ou plusieurs des caractéristiques suivantes : la fonction-objectif (minimiser la distance intra-classe ou maximiser la distance inter-classe), la méthode de normalisation (distance à 1 ou somme des poids égale à 1), les variables souples (souplesses sur les distances intra- et/ou inter-classes) et la symétrie de ces variables souples (souplesses en i et j, ou en i seulement, pour la distance $d_{i,j}$). Par exemple, le modèle $Max{-}w{-}\xi^E_{i+j}$ maximise la distance inter-classe (Max), normalise avec la somme des poids égale à 1 (w), autorise la souplesse pour les distances inter-classes (ξ^E) avec des variables différentes (ξ_{i+j}) ; il s'écrit :

$$\begin{array}{lllll} \max & d_{inter} - C\sum_{i=1}^{N} \xi_i^{inter} & & & \\ s.t. & \sum_{m=1}^{M} w_m a_{i,j,m} & \geq & d_{inter} - \xi_i^{inter} - \xi_j^{inter} & (y_i \neq y_j) \\ & \sum_{m=1}^{M} w_m & = & 1 & \end{array}$$

En combinant les différentes caractéristiques, nous avons ainsi proposé un total de 40 modèles différents avec des performances variables [3].

Dans cette présentation, nous avons évalué ces modèles et analyser leurs performances par une étude expérimentale. Nous avons essayé de conclure et de trouver les bons modèles et/ou les bonnes caractéristiques selon les circonstances.

2 Étude expérimentale

Chaque modèle est évalué sur les mêmes jeux de données et selon le même protocole que celui proposé par Hocke et Martinez [2] : une validation croisée de 10 fois 5-folds. Les deux critères d'évaluation sont le taux d'erreur et le temps de calcul. Pour mettre en évidence d'éventuelles différences de performances significatives, nous utilisons différents outils statistiques ainsi que les méthodes de comparaisons recommandées par Demšar [1].

Le premier constat, clair, est qu'il existe des différences significatives entre certains modèles. De plus, les différentes caractéristiques n'ont pas toutes la même influence.

Pour la fonction-objectif, minimiser la distance intra-classe ou maximiser la distance inter-classe ne fait guère de différence. Par contre, la normalisation par les poids est très nettement et systématiquement dominée par la normalisation par la distance à 1.

Intégrer de la souplesse permet d'améliorer les prédictions, mais au détriment du temps de calcul. Plus précisément : sans aucune souplesse, les temps de calcul sont faibles (pas de paramètre C à déterminer), mais le taux d'erreur peut être amélioré. Ajouter un degré de souplesse dégrade le temps de calcul (11 modèles à résoudre pour déterminer C), mais améliore parfois significativement le taux d'erreur ; une souplesse sur les distances intra-classe est moins efficace qu'une souplesse sur les distances inter-classes. Permettre une souplesse sur les deux distances dégrade beaucoup les temps de calculs (121 modèles à résoudre) pour un taux d'erreur peu amélioré. Enfin, permettre une symétrie des variables souples dégrade légèrement les temps de calcul pour un gain faible sur les taux d'erreur.

Au final, cette étude permet d'une part d'identifier des modèles qui sont toujours dominés et qui peuvent donc être ignorés, et d'autre part de mettre en avant des modèles permettant un compromis entre taux d'erreur et temps de calcul intéressant. Le choix du «bon» modèle dépend alors du cas d'usage.

À noter que si on utilise des jeux de données normalisés, les résultats obtenus sont un peu différents (taux d'erreurs et temps de calcul légèrement améliorés), mais cela ne remet pas en cause la comparaison entre les différents modèles.

La première limite ou perspective de ce travail vient du calcul de la valeur des paramètres C : la méthode utilisée, qui essaie 11 valeurs candidates, peut sans doute être grandement améliorée. Il semble en effet possible d'éliminer certaines valeurs et de gagner ainsi beaucoup de temps. Cela permettrait d'utiliser les modèles avec les meilleurs taux d'erreur sans en payer le coût de calcul actuel.

Une deuxième perspective est de travailler à la sélection d'observations et/ou d'attributs pour permettre un meilleur passage à l'échelle et améliorer les performances.

Enfin, avoir des poids optimisés pour chaque variable pourrait être interprété afin de faire du sens et d'extraire de la connaissance sur le problème traité.

Références

[1] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7 :1–30, 2006.

[2] Jens Hocke and Thomas Martinetz. Maximum distance minimization for feature weighting. *Pattern Recognition Letters*, 52 :48–52, 2015.

[3] Yuzhen Wang, Pierre Lemaire, Iragaël Joly, and Nadia Brauner. Programmation linéaire pour améliorer la performance de K-NN avec la distance euclidienne pondérée. In *ROADEF 2021*, Mulhouse, France, April 2021.

[4] Ian H Witten and Eibe Frank. Data mining : practical machine learning tools and techniques. *Morgan Kaufmann*, 2005.