Algorithme des K plus proches voisins pondérés (WKNN)

et

Application en diagnostic

Eve MATHIEU-DUPAS
Responsable Plate-forme bioinformatique & biostatistique
UMR SysDiag, Modélisation et ingénierie des systèmes complexes biologiques pour le diagnostique

eve.dupas@sysdiag.cnrs.fr
SysDiag, Unité Mixte de Recherche CNRS-BIO-RAD
1682 Rue de la valsière, CS61003
34184 Montpellier Cedex 4

Résumé : La méthode des k plus proches voisins pondérés est une méthode de classification supervisée offrant des performances très intéressantes dans la recherche de nouveaux biomarqueurs pour le diagnostic. Nous présentons le fondement théorique de cette méthode et illustrerons cette technique d'apprentissage statistique au travers du problème diagnostique d'une pathologie complexe.

Summary : The Weighted k-nearest neighbor method is a supervised classification approach which provides very interesting results in the identification of new biomarkers for the diagnostic research. The theory is shown and illustrated through a complex pathology.

Mots-clés : Apprentissage statistique. Clasiification supervisée. Méthode à base de voisinage. Diagnostic et découverte de biomarqueurs.

1- Introduction

Les méthodes d'apprentissage statistique et de classification [1] sont d'un intérêt majeur en recherche diagnostique, plus précisément dans l'identification des combinaisons de biomarqueurs qui constitueront les futurs tests de diagnostic in vitro. La méthode des k plus proches voisins pondérés [2], figurant parmi les méthodes à base de voisinage, offre dans ce contexte des performances très intéressantes.

2- Méthodologie

2.1- L'algorithme KNN

L'algorithme KNN figure parmi les plus simples algorithmes d'apprentissage artificiel. Dans un contexte de classification d'une nouvelle observation $\boldsymbol{x}$, l'idée fondatrice simple est de faire voter les plus proches voisins de cette observation. La classe de $\boldsymbol{x}$ est déterminée en fonction de la classe majoritaire parmi les k plus proches voisins de l'observation $\boldsymbol{x}$.
La méthode KNN est donc une méthode à base de voisinage, non-paramétrique ; Ceci signifiant que l'algorithme permet de faire une classification sans faire d'hypothèse sur la fonction $y=f(x_1,x_2,\ldots x_p)$ qui relie la variable dépendante aux variables indépendantes.

Algorithme 1-NN

La méthode du plus proche voisin est une méthode non paramétrique où une nouvelle observation est classée dans la classe d'appartenance de l'observation de l'échantillon d'apprentissage qui lui est la plus proche, au regard des covariables utilisées. La détermination de leur similarité est basée sur des mesures de distance.
Formellement, soit *L* l'ensemble de données à disposition ou échantillon d'apprentissage :

$$L = \{(y_i, x_i), i = 1,..., n_L\}$$

où $y_i \in \{1,...,c\}$ dénote la classe de l'individu *i* et le vecteur $\mathbf{x_i}' = (x_{i1},..., x_{ip})$ représente les variables prédictrices de l'individu *i*. La détermination du plus proche voisin est basée sur un fonction distance arbitraire *d(.,.)* .
La distance euclidienne ou dissimilarité entre deux individus caractérisés par *p* covariables est définie par:

$$d((x_1, x_2,..., x_p),(u_1, u_2,...,u_p)) = \sqrt{(x_1-u_1)^2 + (x_2 - u_2)^2 + ... + (x_p - u_p)^2}$$

Ainsi, pour une nouvelle observation *(y,**x**)* le plus proche voisin *($y_{(1)}$,$\mathbf{x_{(1)}}$)* dans l'échantillon d'apprentissage est déterminé par :

$$d(\mathbf{x}, \mathbf{x_{(1)}}) = \min_i (d(\mathbf{x}, \mathbf{x_i}))$$

et $\hat{y} = y_{(1)}$, la classe du plus proche voisin, est sélectionnée pour la prédiction de *y*. Les notations $x_{(j)}$ et $y_{(j)}$ représentent respectivement le $j^{ème}$ plus proche voisin de *x* et sa classe d'appartenance.

Parmi les fonctions distance types, la distance euclidienne est définie comme suit :

$$d(\mathbf{x_i}, \mathbf{x_j}) = \left(\sum_{s=1}^{p} (x_{is} - x_{js})^2 \right)^{\frac{1}{2}}$$

et plus généralement la distance de Minkowski :

$$d(\mathbf{x_i}, \mathbf{x_j}) = \left(\sum_{s=1}^{p} |x_{is} - x_{js}|^q \right)^{\frac{1}{q}}$$

La méthode est justifiée par l'occurrence aléatoire de l'échantillon d'apprentissage. La classe $Y_{(1)}$ du voisin le plus proche $\mathbf{x_{(1)}}$ d'un nouveau cas $\mathbf{x}$ est une variable aléatoire. Ainsi la probabilité de classification de $\mathbf{x}$ dans la classe $y_{(1)}$ est P[$Y_{(1)}$ / $\mathbf{x_{(1)}}$]. Pour de grands échantillons d'apprentissage, les individus $\mathbf{x}$ et $\mathbf{x_{(1)}}$ coincident de très près, si bien que $P[y_{(1)} / \mathbf{x_{(1)}}] \approx P[y / \mathbf{x}]$. Ainsi, la nouvelle observation (individu) $\mathbf{x}$ est prédite comme appartenant à la vraie classe *y* avec une probabilité égale approximativement à P[*y*/$\mathbf{x}$].

Algorithme KNN

Une première extension de cette idée, qui est largement et communément utilisée en pratique, est la méthode des k plus proches voisins. La plus proche observation n'est plus la seule observation utilisée pour la classification. Nous utilisons désormais les *k* plus proches observations. Ainsi la décision est en faveur de la classe majoritairement représentée par les *k* voisins. Soit k_r le nombre d'observations issues du groupe des plus proches voisins appartenant à la classe *r*

$$\sum_{r=1}^{c} k_r = k$$

Ainsi une nouvelle observation est prédite dans la classe *l* avec :

$$l = \max_r (k_r)$$

Ceci évite que la classe prédite ne soit déterminée seulement à partir d'une seule observation. Le degré de localité de cette technique est déterminé par le paramètre *k* : pour *k*=1, on utilise la méthode du seul plus proche voisin comme technique locale maximale, pour $k \to n_l$ on utilise la classe majoritaire sur l'ensemble intégral des observations (ceci impliquant une prédiction constante pour chaque nouvelle observation à classifier).

Quelques règles sur le choix de k **:**
Le paramètre *k* doit être déterminé par l'utilisateur : $k \in N$. En classification binaire, il est utile de choisir *k* impair pour éviter les votes égalitaires. Le meilleur choix de k dépend du jeu de donnée.En général, les grandes valeurs de *k* réduisent l'effet du bruit sur la classification et donc le risque de sur-apprentissage, mais rendent les frontières entre classes moins distinctes. Il convient donc de faire un choix de compromis entre la variabilité associée à une faible valeur de *k* contre un 'oversmoothing' ou surlissage (i.e gommage des détails) pour une forte valeur de *k*. Un bon *k* peut être sélectionné par diverses techniques heuristiques, par exemple, de validation-croisée. Nous choisirons la valeur de k qui minime l'erreur de classification.

2.2 - Méthode des k plus proches voisins pondérés et classification ordinale

Similarité entre voisins

Cette technique étend la méthode des k plus proches voisins selon deux voies :

1) Tout d'abord, un schéma de pondération des plus proches voisins est introduit en fonction de leur similarité avec la nouvelle observation à classer
2) Basé sur le fait que le vote des plus proches voisins est équivalent au mode de la distribution de la classe, la seconde extension utilise la médiane ou la moyenne de cette distribution, si la variable cible est relative à une échelle ordinale ou de niveau plus élevé.

Cette extension est fondée sur l'idée que les observations de l'échantillon d'apprentissage, qui sont particulièrement proches de la nouvelle observation (y,$\boldsymbol{x}$), doivent avoir un poids plus élevé dans la décision que les voisins qui sont plus éloignés du couple (y,$\boldsymbol{x}$).
Ce n'est pas le cas avec la méthode KNN : en effet seuls les k plus proches voisins influencent la prédiction, mais l'influence est identique pour chacun des voisins, indépendamment de leur degré de similarité avec (y,$\boldsymbol{x}$). Pour atteindre ce but, les distances, sur lesquelles la recherche des voisins est fondée dans une première étape, sont transformées en mesures de similarité, qui peuvent être utilisées comme poids.

Standardisation des covariables (afin d'homogénéiser le calcul des distances)

Dans une première étape, les *k* plus proches voisins sont sélectionnés selon la distance de Minkowski, en supposant que les deux paramètres *k* et *q* aient été fixés par l'utilisateur.

Afin de pondérer de façon équitable chacune des covariables pour le calcul des distances, les valeurs doivent être standardisées. Dans un contexte de ratio ou de différence, cet objectif est atteint

en divisant simplement les variables par leur déviation standard.

Système de pondération pour le voisinage : la fonction Noyau

La transition de distances en poids se fait dans une deuxième étape selon une fonction noyau arbitraire. Les noyaux sont des fonctions *K(.)* maximales en *d*=0 et décroissantes avec des valeurs de *d* grandissantes en valeur absolue. La fonction noyau *K(x)* est généralement symétrique et doit être telle que

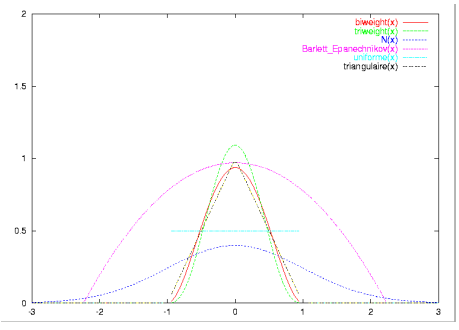
$$\int_{-\infty}^{+\infty} K(x)dx = 1$$

Elle doit vérifier les propriétés suivantes :

- $K(d) \geq 0$ pour tout $d \in \Re$
- $K(d)$ atteint son maximum pour $d=0$
- $K(d)$ décroissante pour $d \to \pm\infty$

Suivent des exemples typiques de fonctions noyaux :

- Rectangulaire (loi uniforme): $\frac{1}{2} I(|d| \leq 1)$
- Triangulaire : $(1-|d|)I(|d| \leq 1)$
- Epanechnikov : $\frac{3}{4}(1-d^2)I(|d| \leq 1)$
- Bi-poids : $\frac{15}{16}(1-d^2)^2 I(|d| \leq 1)$
- Tri-poids : $\frac{35}{32}(1-d^2)^3 I(|d| \leq 1)$
- Cosine $\frac{\pi}{4}\cos\left(\frac{\pi}{2}d\right)I(|d| \leq 1)$
- *gaussien* : $\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}d^2)$
- Inverse $\frac{1}{|d|}$
- *Barlett – Epanechnikov* : $\frac{3}{4}\frac{\left(1-\frac{d^2}{5}\right)}{\sqrt{5}}$ si $|d| < \sqrt{5}, 0$ sinon



Dans le cas de distances, qui sont définies comme fonctions positives, bien sûr seul le domaine positif de K peut être utilisé. En ce sens, le choix du noyau est le troisième paramètre de cette technique. Mais, le choix de ce noyau (excepté pour le cas rectangulaire, où tous les poids sont égaux) ne se révèle pas être crucial dans l'obtention des résultats.
Le but de la fonction noyau est de pondérer les observations par rapport à un point de référence de sorte que plus une observation est proche de la référence, plus son poids sera important. On donne ainsi plus de poids aux observations proches de celle qu'on cherche à estimer qu'aux autres observations.

Chaque fonction noyau nécessite soit une largeur de fenêtre, si les valeurs deviennent nulles à partir d'une certaine distance de la valeur maximale, soit un paramètre de dispersion si les valeurs sont strictement positives pour tout $d \in \Re$. Dans la procédure KKNN, les deux sont sélectionnés automatiquement en fonction de la distance du premier voisin $x_{(k+1)}$ qui n'est plus pris en considération. Ceci est fait implicitement en standardisant tous les autres distances par la distance du $(k+1)^{\text{ème}}$ voisin :

$$D(x, x_{(i)}) = \frac{d(x, x_{(i)})}{d(x, x_{(k+1)})} \text{ pour } i = 1,...,k$$

Ces distances standardisées prennent ainsi toujours des valeurs dans l'intervalle [0,1]. Dans notre implémentation, nous ajoutons une constante $\varepsilon>0$ à $d(x,x_{(k+1)})$ pour éviter des poids de 0 pour certaines des plus proches voisins.

Règle de classification d'une nouvelle observation
Après détermination des mesures de similarités entre observations, chaque nouveau cas $(y,\boldsymbol{x})$ est attribué à la classe l de poids maximal, dans son voisinage à k voisins :

$$l = \max_r \left(\sum_{i=1}^{k} \underbrace{K\left(D\left(x, x_{(i)}\right)\right)} I\left(y_{(i)} = r\right) \right)$$

Poids cumulé des voisins parmi les kNN qui appartiennent à la classe r

Les algorithmes KNN et 1-NN peuvent être vus comme cas particuliers : noyau rectangulaire pour KNN , et k=1 indépendamment du noyau choisi pour 1-NN

Le principal objectif de cette extension de méthode est d'obtenir une méthode qui jusqu'à un certain degré est indépendante d'un mauvais choix de k générant un taux élevé d'erreur de classification. Si k est choisi trop grand, les poids réduisent l'influence des voisins qui sont trop éloignés de la nouvelle observation.

Les étapes de la classification wkNN

1. Soit L un échantillon d'apprentissage constitué des observations $\mathbf{x_i}$ relatives à une classe y_i :
$$L = \{(y_i, \mathbf{x_i}), i = 1,..., n_L\}$$

Soit **x** une nouvelle observation, dont la classe y doit être prédite :

$$\hat{y} = ?$$

2. sélection des *(k+1)* plus proches voisins de ***x*** selon une fonction distance d(.,.) préalablement choisie :

$$d(x,x_i)$$

3. Standardisation des k plus petites distances via le $(k+1)^{\text{ème}}$ voisin :

$$D_{(i)} = D\left(\mathbf{x}, \mathbf{x_{(i)}}\right) = \frac{d(\mathbf{x}, \mathbf{x_{(i)}})}{d(\mathbf{x}, \mathbf{x_{(k+1)}})}$$

4. Transformation des distances normalisées $D_{(i)}$ en poids $w_{(i)}$ à partir d'une fonction noyau K(.) :

$$w_{(i)} = \mathrm{K}(D_{(i)})$$

5. La classe de *x* est choisie d'après la majorité pondérée des *k* plus proches voisins :

$$\hat{y} = \max_r \left(\sum_{i=1}^{k} w_{(i)} I\left(y_{(i)} = r\right) \right)$$

Evaluation de la méthode : taux d'erreur ou de mauvaise classification

L'évaluation de la méthode KKNN est basée sur le taux d'erreur de classification :

$$\tau = \frac{1}{n} \sum_{i=1}^{n} I\left(y_i \neq \hat{y}_i\right)$$

Nous pourrions montrer que si on disposait d'un très gros volume de données d'apprentissage et en utilisant une règle de classification arbitrairement sophistiquée, nous ne diminuerions l'erreur de mauvaise classification que d'un facteur 2 par rapport à une méthode 1-NN

$$\tau_{erreur-1NN} \leq 2 \times \tau_{erreur-Bayes}$$

Sélection échantillons d'apprentissage et tests :
L'ensemble des données est divisé aléatoirement en deux parties, consistant respectivement en les 2/3 et 1/3 des données. L'échantillon d'apprentissage est utilisé comme un ensemble de prototypes et les observations de l'échantillon test sont prédites. Nous utilisons 50 tirages aléatoires d'échantillons d'apprentissages et tests et calculons la moyenne des taux d'erreurs sur ces 50 tirages.

3- Application

La méthode des k plus proches voisins pondérés sera illustrée dans le cadre de la recherche de nouveaux biomarqueurs pour le diagnostic d'une pathologie complexe.

Bibliographie

[1] Hastie, T. & al. (2001) *The Elements of Statistical Learnng*, Springer, Canada.
[2] Hechenbichler, K. Et Schliep K. (2004) Weighted k-nearest-neighbor techniques and ordinal classification. *Sonderforschungsbereich 386, paper 399.*.