

Application des méthodes d'apprentissage pour l'étude des risques de pollution dans le Lac Léman.

**Nicolas GILARDI^{1,2}, Alex GAMMERMAN³, Mikhail KANEVSKI^{2,3}, Michel MAIGNAN¹,
Tom MELLUISH³, Craig SAUNDERS³, Volodia VOVK³**

¹ : Institut de Minéralogie, Université de Lausanne, Lausanne, Suisse

² : Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP), Martigny, Suisse

³ : Department of Computer Science, Royal Holloway, University of London, Egham, UK

⁴ : Institute of Nuclear Safety (IBRAE), Moscow, Russia

Introduction

Depuis quelques années, de nouvelles méthodes d'apprentissage se développent sur la base de la Théorie de l'Apprentissage Statistique (Statistical Learning Theory) de Vapnik et Chervonenkis [Vapnik, 1995]. L'une de ces méthodes, appelée Machine à Vecteur de Support ou SVM (Support Vector Machine) [Cortes et al., 1995], permet de réaliser des estimations en classification (à deux classes ou plus) [Burges, 1998] ou en régression [Smola et al., 1998].

De telles méthodes permettent généralement de s'affranchir de contraintes statistiques sur les données étudiées comme la normalité de la distribution. De plus, elles sont non linéaires ce qui leur donne un pouvoir de généralisation supérieur dans certains cas, aux méthodes de régressions plus classiques.

Cependant, ces méthodes, comme beaucoup d'autres, ne permettent pas d'obtenir d'intervalle de confiance sur l'estimation effectuée. Ce problème est résolu par les méthodes de « transduction universelle » développées à l'Université du Royal Holloway [Gammerman et al., 1998]. Sur la base d'une méthode de classification ou de régression quelconque, ces méthodes permettent d'établir une probabilité de dépasser une valeur donnée.

Dans ce document, nous présenterons tout d'abord une théorie de la classification avec les SVM puis les résultats de l'utilisation de cette méthode pour l'étude de la pollution des sédiments du Lac Léman. Ensuite, nous expliquerons les grandes lignes du principe de la méthode de transduction développée par les chercheurs du Royal Holloway. Et en conclusion, nous montrerons comment ces deux méthodes peuvent, de par leur principe même, se révéler complémentaires dans le développement de cartes de risque.

Théorie des Support Vector Machines

Les Support Vector Machines sont une classe d'algorithmes basés sur le principe de minimisation du « risque structurel » décrit par la Théorie de l'Apprentissage Statistique de Vapnik et Chervonenkis [Vapnik, 1995] [Schölkopf et al., 1999].

Minimisation du risque structurel

Lorsque l'on utilise des méthodes d'apprentissage, on utilise généralement deux jeux de données principaux : le jeu d'entraînement et le jeu de test. Le jeu d'entraînement représente la part des données d'origine utilisée pour calculer le modèle, et le jeu de test est l'autre partie, inconnue de l'algorithme d'apprentissage et utilisée pour évaluer les performances de généralisation du modèle. La qualité de ce modèle est alors jugée à sa capacité à réduire l'erreur de test ou de « généralisation ».

Cependant, comme le modèle n'est pas construit en utilisant le jeu de test, l'erreur de généralisation ne peut pas être évaluée exactement car elle dépend de la distribution de probabilité des données :

$$R[f] = \int \frac{1}{2} Q(\mathbf{x}) dP(\mathbf{x}, y)$$

où Q est la fonction d'erreur (erreur absolue dans le cas des SVM), \mathbf{x} est le vecteur d'entrée, et $P(x,y)$ est la distribution des données (qui nous est inconnue).

La seule information dont nous disposons comme évaluation de l'erreur est l'erreur d'entraînement :

$$R_{emp}[f] = \frac{1}{l} \sum_{i=1}^l \frac{1}{2} Q(\mathbf{x}_i)$$

où l est le nombre de points d'entraînement.

Ce n'est pas suffisant. La raison en est que l'on peut facilement trouver un modèle minimisant l'erreur d'entraînement mais pour lequel l'erreur de généralisation sera très grande. Un exemple simple est la régression de données linéaires bruitées au moyen d'une fonction polynomiale : plus le degré du polynôme sera grand, plus l'erreur d'entraînement sera faible, mais plus l'erreur de généralisation sera élevée. On peut donc comprendre que cette dernière est aussi liée à la famille de fonction utilisée comme modèle. Cette dépendance est nommée « risque structurel ».

Dans leur Théorie de l'Apprentissage Statistique, Vapnik et Chervonenkis ont prouvé qu'il est possible de définir une majoration du risque structurel en fonction de la famille de fonction utilisée pour le modèle. L'une de ces majorations peut être calculée en utilisant la dimension de Vapnik-Chervonenkis (dimension VC) qui représente le plus grand nombre de points pouvant être séparés de toutes les façons possibles par un membre de l'ensemble de fonction. La borne VC est alors définie ainsi : si la dimension VC h de la famille de fonction utilisée est inférieure au nombre de points d'entraînement l , alors avec une probabilité d'au moins $1 - \eta$, on a :

$$R[f] \leq R_{emp}[f] + \sqrt{\frac{h \left(\log \left(\frac{2l}{h} \right) + 1 \right) - \log \left(\frac{\eta}{4} \right)}{l}}$$

Classification par hyperplan

Considérons maintenant l points $x_i \in R^N$, $i = 1, \dots, l$, séparés en deux classes définies par $f(x_i) = -1$ et $f(x_j) = 1$, linéairement séparables dans R^N . Classons ces points en utilisant une famille de fonctions linéaires définie par $(w \cdot x) + b = 0$, avec $w \in R^N$, $b \in R$, de sorte que $f(x) = \text{sgn}((w \cdot x) + b)$. Il a été démontré que maximiser la distance (la marge) entre un hyperplan de cette famille de fonction et chacune des deux classes de points réduit la dimension VC de la famille de fonction décrivant le classifieur. De plus, il existe un seul hyperplan pour lequel la marge est maximum.

Donc pour trouver le meilleur classifieur, en terme de minimisation du risque structurel, il faut résoudre le problème d'optimisation sous contrainte suivant :

$$\text{minimiser} \quad \tau(w) = \frac{1}{2} \|w\|^2$$

$$\text{sachant que} \quad y_i \cdot ((w \cdot x_i) + b) \geq 1, \quad f(x_i) = y_i, \quad i = 1, \dots, l$$

Algorithme des Support Vectors

Ce genre de problème d'optimisation se résout en introduisant des multiplicateurs de Lagrange $\alpha_i \geq 0$ et le Lagrangien

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^l \alpha_i (y_i \cdot ((x_i \cdot w) + b) - 1)$$

D'autres contraintes sont nécessaires pour résoudre ce problème car le Lagrangien doit être minimisé selon w et b et maximisé selon α . Ceci conduit à une formulation duale de notre problème, plus simple, qui est l'expression de l'algorithme des Support Vectors pour des données linéairement séparables :

$$\text{maximiser} \quad W(\alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

$$\text{sachant que} \quad \alpha_i \geq 0, \quad i = 1, \dots, l, \quad \text{and} \quad \sum_{i=1}^l \alpha_i y_i = 0$$

la fonction de décision est alors :

$$f(x) = \text{sgn} \left(\sum_{i=1}^l (y_i \alpha_i (x \cdot x_i)) + b \right)$$

Cette fonction de décision est donc seulement influencée par les points correspondants à des valeurs non nulles de α_i . Ces points sont appelés les Vecteurs de Support (Support Vectors). Ils correspondent, dans un cas linéairement séparable, aux points les plus proches de la limite de décision, c'est à dire aux points se trouvant exactement sur la marge. Il s'agit là d'une propriété très intéressante des SVM : seuls les Support Vectors sont nécessaires pour décrire cette limite de décision, et le nombre de SV pour le modèle optimal est généralement petit devant le nombre de données d'entraînement.

Généralisation

Bien sûr, il est assez rare d'avoir des données linéairement séparables. Afin de traiter également des données bruitées ou non linéairement séparables, les SVM ont été généralisées grâce à deux outils : la « marge souple » (soft margin) et les « fonctions noyau » (kernel functions).

Le principe de la marge souple est d'autoriser des erreurs de classification. La nouvelle formulation du problème d'optimisation est alors :

$$\text{minimiser} \quad \tau(w, \xi) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i, \quad C > 0$$

$$\text{sachant que} \quad y_i \cdot ((w \cdot x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad f(x_i) = y_i, \quad i = 1, \dots, l$$

Le paramètre C est défini par l'utilisateur. Il peut être interprété comme une tolérance au bruit du classifieur : pour de grandes valeurs de C, seules de très faibles valeurs de ξ sont autorisées, et par conséquent, le nombre de points mal classés sera très faible (données faiblement bruitées). *A contrario*, si C est petit, ξ peut devenir très grand, et on autorise alors bien plus d'erreur de classification (données fortement bruitées). Le nouvel algorithme est donc :

$$\text{maximiser} \quad W(\alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

$$\text{sachant que} \quad 0 \leq \alpha_i \leq C, \quad i = 1, \dots, l, \quad \text{and} \quad \sum_{i=1}^l \alpha_i y_i = 0$$

La seule différence avec le cas linéairement séparable est donc l'introduction d'une borne supérieure pour les paramètres α . Il est également intéressant de noter que les points se trouvant, par l'introduction de cette technique, du « mauvais » côté de la limite de décision sont tous des Support Vectors, quelle que soit leur distance à cette limite, ce qui signifie qu'ils exercent une influence sur le calcul de cette limite.

Maintenant, que faire si les données ne sont pas linéairement séparables ? L'idée est de projeter notre espace d'entrée dans un espace de plus grande dimension afin d'obtenir une configuration linéairement séparable (à l'approximation de la marge souple près) de nos données, et d'appliquer alors l'algorithme initial des SVM. Le nouvel algorithme peut donc être écrit ainsi :

$$\text{partant de} \quad \Phi : R^N \rightarrow R^M, \quad M > N$$

$$\text{maximiser} \quad W(\alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j (\Phi(x_i) \cdot \Phi(x_j))$$

$$\text{sachant que} \quad 0 \leq \alpha_i \leq C, \quad i = 1, \dots, l, \quad \text{and} \quad \sum_{i=1}^l \alpha_i y_i = 0$$

Mais ce genre de transformation peut devenir très coûteux du point de vue calcul pour de grandes valeurs de M. Afin d'éviter ce problème, l'idée est d'utiliser les fonctions noyau de Mercer. La

principale propriété de ces fonctions est que, partant d'une projection $\Phi : R^N \rightarrow R^M$ et deux vecteurs $x, y \in R^N$, on a $k(x, y) = \Phi(x) \cdot \Phi(y)$. De ce fait, il n'est pas nécessaire de procéder à cette projection : la fonction noyau donne le même résultat dans l'espace d'entrée. Une difficulté possible reste de choisir une fonction noyau efficace.

Nous avons maintenant un algorithme général pour résoudre n'importe quel problème de classification :

maximiser
$$W(\alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j k(x_i \cdot x_j)$$

sachant que $0 \leq \alpha_i \leq C, \quad i = 1, \dots, l, \quad \text{and} \quad \sum_{i=1}^l \alpha_i y_i = 0$

et la nouvelle fonction de décision est alors :

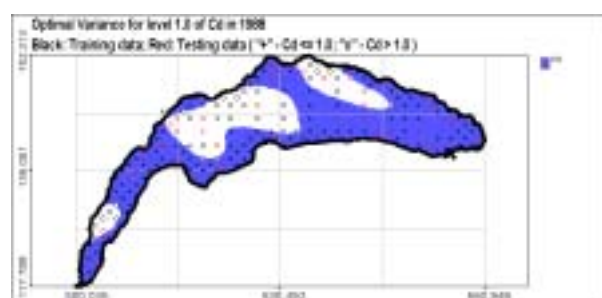
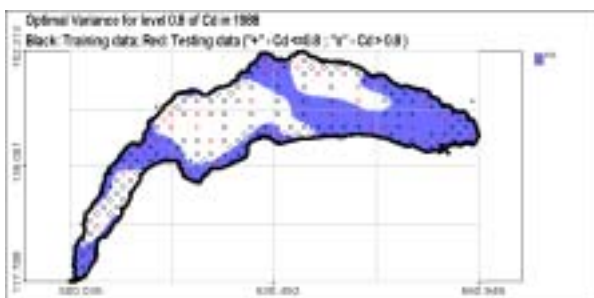
$$f(x) = \text{sgn} \left(\sum_{i=1}^l (y_i \alpha_i k(x \cdot x_i)) + b \right)$$

Application à des études de pollution

Afin d'évaluer l'efficacité de ces méthodes pour traiter des données spatiales, nous les avons comparées à des techniques plus classiques, comme le krigeage des indicatrices [Gilardi et al., 1999]. Nous avons pour cela utilisé un jeu de données fourni par la CIPEL (Commission Internationale pour la Protection des Eaux du Léman). Il s'agit d'une série d'analyses physico-chimiques effectuées sur les sédiments du Lac Léman au cours des années 1978, 1983 et 1988. Pour nos expériences, nous avons utilisé les analyses de 1988 de concentration en Cadmium.

Les données de la CIPEL étant continues, nous avons défini nos classes selon le principe de la valeur seuil : au-dessus d'une certaine teneur, nous considérons que le point appartient à la classe -1, en dessous, il appartient à la classe +1. Cette méthode nous permet donc de dresser une carte des zones du Lac dépassant une certaine concentration de cadmium dans leurs sédiments.

Les teneurs utilisées, en microgrammes de cadmium par grammes de sédiments, sont de 0,8 et 1,0. Le choix de ces valeurs, proche de la concentration moyenne, n'a aucune origine légale ou sanitaire. Il est avant tout motivé par le désir d'avoir un nombre suffisant de représentants des deux classes.



Classifications optimales pour les niveaux 0.8 ppm et 1.0 ppm de Cd

Les résultats de ces classifications se sont révélés conformes à ceux de la méthode par krigeage des indicatrices et ont montré un bon potentiel des SVM pour traiter des données spatiales.

Mais si, dans le cas d'une étude environnementale, il est intéressant de posséder une méthode qui fournit une bonne évaluation des zones polluées, il est encore plus intéressant de posséder une information de « confiance » sur cette évaluation. Ceci permet d'éviter que les quelques erreurs faites par la méthode ne puissent avoir de graves conséquences.

C'est ce que se propose de faire la méthode dite de « p-value Transduction » développée au département d'informatique de l'université de Royal Holloway [Gammerman et al., 1998]

Théorie de la « p-value Transduction »

Selon la définition de Vapnik [Vapnik,1995], le terme de *transduction* s'applique à des méthodes d'apprentissage s'attachant à résoudre, à l'aide de données particulières, un problème particulier. Ceci s'oppose à la méthode *inductive* qui, toujours à partir de données particulières, produit un résultat général (un modèle) et l'applique ensuite de façon identique à toute nouvelle donnée. Dans l'absolu, une méthode transductive construira autant de modèles qu'il y aura de points à estimer. L'application de ce principe a donné lieu à de multiples applications ayant montré d'excellents résultats en terme de généralisation.

La méthode présentée ici s'intéresse d'avantage à l'évaluation d'un intervalle de confiance et d'une mesure de crédibilité du résultat qu'à la qualité du résultat lui-même. La méthode présentée ci-dessous évalue ces intervalles à partir d'une classification par SVM [Saunders et al., 1999]. D'autres travaux ont montré qu'il était possible d'appliquer ce même principe à d'autres types de classifieur, mais aussi à la régression¹.

Principe

Comme montré dans le paragraphe « Algorithme des Support Vectors », la méthode des SVM consiste à résoudre un problème d'optimisation quadratique dont les solutions sont les multiplicateurs de Lagrange α . La valeur de ces multiplicateurs indique le poids relatif des points de l'ensemble d'entraînement dans la construction de la limite de décision entre les classes : les points pour lesquels ces multiplicateurs sont nuls n'interviennent pas du tout dans la construction de cette limite.

Par conséquent, l'idée de dire que ces paramètres représentent, d'une certaine façon, « l'étrangeté » d'un point paraît appropriée : plus son multiplicateur de Lagrange est grand et moins il est représentatif de la classe à laquelle il appartient².

C'est cette mesure d'étrangeté qui va nous permettre de construire nos intervalles de confiance.

¹ Des travaux sur ce sujet sont menés par Tom Melliush : thomasm@dcs.rhnc.ac.uk.

² Il n'y a pas de paradoxe à cela : l'algorithme des SVM modélise la limite entre les classes et non leurs zones d'influence. Par conséquent ce sont les points « limites » qui servent de base à la modélisation et non les points représentatifs.

Définitions des p-values

Supposons maintenant qu'à partir de l points d'entraînement répartis en 2 classes, nous souhaitons déterminer la classe à laquelle appartient un nouveau point mais également la confiance que nous pouvons accorder à notre résultat. Nous allons donc construire deux modèles de SVM : l'un dans lequel nous supposons que notre nouveau point appartient à la classe -1, et l'autre supposant qu'il appartient à la classe +1. De ces deux modèles, nous allons obtenir pour notre point deux valeurs du multiplicateur de Lagrange correspondant α_n .

Si l'on suppose que nos données sont identiquement et indépendamment distribuées (iid), alors on peut

$$P\left\{\alpha_n > \max_{1 \leq i \leq l} \alpha_i\right\} \leq \frac{1}{l+1}$$

dire que la probabilité que α_n ait la plus grande valeur parmi tous les α_i du modèle étudié est :

Les p-values correspondant à notre nouveau point sont définies ainsi: connaissant le rang n de la valeur α_n au sein de l'ensemble des valeurs α_i , la p-value associée est la probabilité que le rang de α_n soit supérieur à n . Autrement dit, la p-value est la probabilité que notre point soit « conforme » à la classe à laquelle il est associé.

$$\text{p-value} = \frac{\text{nbr}\{i : \alpha_i \geq \alpha_n\}}{l+1}$$

Cette probabilité peut s'écrire comme suit, toujours en supposant nos données iid:

Dans le cas de notre exemple, nous allons obtenir deux p-values: une pour chaque classification. Il est bien évident que l'une de ces deux classifications est mauvaise, notre point appartenant à une et seulement une des deux classes. C'est la plus grande des valeurs des p-values qui va donc nous donner la classe que notre algorithme de SVM a estimé être la bonne.

Confiance et crédibilité

Maintenant, qu'en est-il de la confiance (probabilité que le résultat soit juste) à accorder à notre classification ? Dans [Gammerman et al., 1998] et [Saunders et al. 1999], il est expliqué que si nous appelons P1 la plus grande des deux p-values, et P2 la plus petite, nous pouvons dire que cette confiance est de $1-P2$. Par conséquent, plus P2 sera faible, moins la probabilité que notre point soit conforme à la « mauvaise » classe sera grande, et plus nous aurons de confiance dans notre résultat.

Une autre propriété intéressante concernant ces p-values, décrite dans ces articles, est que l'on peut également fournir une information de crédibilité de nos données. En effet, si P1 est relativement faible (bien que supérieure à P2), un doute peut subsister sur la qualité de notre classification. C'est pourquoi la crédibilité du point traité est définie par la valeur $1-P1$.

Il est bien évident que toutes ces mesures de probabilités reposent sur nos données d'entraînement. Il est donc très important qu'elles soient bien représentatives du phénomène étudié, ou tout du moins de ce que l'on connaît de lui.

Conclusion

Le développement des méthodes d'apprentissage sur la base de la Théorie de l'Apprentissage Statistique a fourni de précieux outils de classification et de régression. L'un d'entre eux, les SVM, s'est révélé tout à fait efficace pour analyser des problèmes environnementaux liés à des phénomènes distribués spatialement.

L'ajout à cette méthode, grâce à une approche transductive, de la possibilité d'évaluer un intervalle de confiance ainsi qu'une information sur la crédibilité des données, ouvre de formidables possibilités pour la construction de cartes de risque. Basée uniquement sur l'information portée par les données, cette méthode en devient plus fiable et plus universelle que des approches imposant une distribution normale ou log-normale. La philosophie transductive est également particulièrement adaptée à l'établissement de carte de risque lorsque l'information désirée ne s'étend pas à toute une région, mais se limite à quelques zones particulières (maisons, zones urbaines, lacs, cours d'eaux, etc...), ce qui est souvent le cas.

Références

- A. Gammerman, V. Vapnik, V. Vovk, *Learning by Transduction, Uncertainty in Artificial Intelligence*, 1998
- C. Saunders, A. Gammerman, V. Vovk, *Transduction with Confidence and Credibility*, 1999.
- B. Schölkopf, C.J.C. Burges, and A.J. Smola. *Advances in kernel methods-Support Vector Learning*. MIT Press, 1999
- C. Cortes and V. Vapnik. *Support vector networks. Machine Learning*, 20: 273–297, 1995.
- V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.
- C.J.C Burges. *A tutorial on Support Vector Machines for patterns recognition. Data Mining and Knowledge Discovery*, 2(2):121-167, 1998.
- A.J. Smola and B. Schölkopf. *A tutorial on Support Vector Regression*, NeuroCOLT2 Technical Report Series, NC2-TR-1998-030. October 1998.
- J. Weston, A. Gammerman, M. Stitson, V. Vapnik, V. Vovk, C. Watkins, *Density Estimation using Support Vector Machines*, Technical Report, Csd-TR-97-23. February 1998.
- N. Gilardi, M. Kanevski, E. Mayoraz, M. Maignan, *Environmental and Pollution Spatial Data Classification with Support Vector Machines and Geostatistics*, 1999.

Mots-clefs

Machines à vecteurs de support, SVM, transduction, pollution, apprentissage, données spatiales, carte de risque.