Modélisation statistique de la toxicité de molécules et domaine de validité : application en chémoinformatique

Jonathan VILLAIN [1,2] & Gilles DURRIEU [1] & Ronan BUREAU [2]

[1] *Laboratoire de Mathématiques de Bretagne Atlantique, Université de Bretagne Sud et UMR CNRS 6205, Campus de Tohannic, 56017 Vannes*
[2] *Centre d'Études et de Recherche sur le Médicament de Normandie, Université de Caen Basse Normandie, Caen*
ronan.bureau@unicaen.fr, {gilles.durrieu,jonathan.villain}@univ-ubs.fr

Résumé. Dans le domaine de la chimie et plus particulièrement en chémoinformatique, des modèles d'estimation des propriétés écotoxicologiques de molécules sont de plus en plus étudiés. Les modèles QSAR (Quantitative Structure Activity Relationship) permettent de prédire le niveau d'activité d'une nouvelle molécule. Cependant, une erreur de prédiction importante du niveau de toxicité des molécules est souvent observée pour des molécules avec un comportement atypique. Nous proposons alors des modèles statistiques robustes permettant de déterminer un domaine de validité et ainsi de déduire la capacité de prédiction d'un modèle pour une molécule.

Mots-clés. classification, détection de nouveauté, écotoxicologie, quantile, régression, SVM.

Abstract. In chemistry and especially in chemoinformatics, models to estimate ecotoxicological properties of chemicals are more and more studied. QSAR models (Quantitative Structure Activity Relationship) predict the level of activity of a new chemicals. However, a large error of prediction for the molecules toxicity is often observed for molecules with atypical behavior. We then propose robust statistical models for determining a domain of validity, and so determine for a molecule the predictive ability of a model.

Keywords. Classification, novelty detection, ecotoxicology, quantile, regression, SVM.

1 Introduction

En chémoinformatique, de nombreux modèles statistiques sont développés pour estimer les propriétés écotoxicologiques de molécules. Parmis ces modèles, les modèles de régression de type QSAR sont souvent utilisés pour décrire la relation entre la structure et l'activité des molécules. Le problème est que le nombre de molécules existant est très important et que les molécules peuvent avoir des modes d'action spécifique envers certains

récepteurs. La mise en place d'un domaine de validité est donc tout aussi important que le modèle lui même pour pouvoir prédire le niveau de toxicité de nouvelles molécules. Afin de pouvoir prédire si une molécule possède un mode d'action spécifique, Neuwoehner et al. (2009) ont introduit le calcul du rapport de toxicité TR. Ce dernier se détermine à partir d'un modèle de régression basé sur la relation entre $\log(1/CE_{50})$ (CE_{50} est la Concentration Efficace médiane) et le coefficient de séparation octanol-eau noté $\log(P)$. Le rapport de toxicité est alors obtenu par le rapport entre les valeurs accessibles dans les bases de données utilisées en chémoinformatique et les valeurs prédites par le modèle de régression. En pratique, les biochimistes considèrent que pour un TR supérieur à 10, la molécule possède un mode d'action spécifique sinon la molécule est considérée comme ayant un mode d'action non-spécifique (toxicité basale). Dans une étude pécédente, Villain et al. (2014) ont établi des modèles de régression quantile robustes sur les données des propriétés physico-chimiques de 401 molécules provenant de différentes base de données accessibles en ligne. Cette modélisation nous a permis de mettre en évidence 3 descripteurs fondamentaux pour définir le mode d'action des molécules qui sont la solubilité moléculaire, le $\log(P)$ et la polarisabilité. Nous voulons prédire, en utilisant ces modèles, le niveau de toxicité de 36 médicaments. Les médicaments sont des molécules qui ont pour but de cibler un récepteur spécifique et ont donc un mode d'action spécifique envers le récepteur ciblé. Dans ce travail, nous étudions des modèles statistiques permettant de déterminer un domaine de validité à partir de ces trois descripteurs afin de préciser si une molécule peut être prédite correctement.

2 Modèlisation et domaine de validité

Dans le domaine de la chémoinformatique, les données sont souvent contaminées par des valeurs extrêmes ou atypiques. Dans les problématiques de prédiction des caractéristiques d'une molécule chimique, il est très difficile de pouvoir prédire correctement les molécules ayant un comportement extrême ou atypique. Des méthodes d'estimation robustes des paramétres des modèles sont alors considérées (Villain et al. (2014a, 2014b)) et une méthode de détection de nouveauté basée sur une approche de classification par SVM (Support Vector Machine) est proposée par Villain et al. (2015).

2.1 Quantiles de régression SVM

Les Support Vector Machines (SVM) ont été développés dans les années 1990 à partir de travaux sur l'apprentissage statistique initiés par Vladimir Vapnik en 1998. On note par $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$, un échantillon statistique de taille n et de distribution inconnue. Dans le cas de régression non linéaire, le principe consiste à rechercher une estimation de $\hat{f}(x)$ d'un modèle $f(x)$ pour Y. Les observations faites dans l'ensemble $\mathcal{F}$ (en général $\mathbb{R}^p$) sont considérées comme étant transformées par une application non linéaire $\boldsymbol{x} \to \phi(\boldsymbol{x})$ qui

va de $\boldsymbol{x}_i \in \mathcal{F}$ pour $i = 1, \ldots, n$ dans un espace muni d'un produit scalaire de plus grande dimension. Nous présentons maintenant la régression non linéaire quantile SVM notée QSMR. La fonction quantile y_i conditionnellement à x_i peut s'écrire pour $i = 1, \ldots, n$:

$$Q(\theta/\boldsymbol{x}_i) = \boldsymbol{w}'_\theta \, \phi(\boldsymbol{x}_i) \qquad \text{pour} \qquad \theta \in (0,1), \tag{1}$$

où $\boldsymbol{w}_\theta$ désigne le θ-quantile de régression. QSVMR peut se définir comme une minimisation pour $\theta \in (0,1)$

$$\frac{1}{2} \, || \, \boldsymbol{w}_\theta \, ||^2 + C \sum_{i=1}^{n} \rho_\theta \left(y_i - \boldsymbol{w}'_\theta \phi(\boldsymbol{x}_i) \right), \tag{2}$$

où C désigne le degré de pénalisation et $\rho_\theta(x) = x \left(\theta - \mathbb{I}(x < 0) \right)$ avec $\mathbb{I}(\mathcal{P})$ qui prend la valeur 1 ou 0 selon que la condition $\mathcal{P}$ est vérifiée ou non. Une solution de (2) pour $\theta \in (0,1)$ s'obtient en optimisant sa version duale quadratique. Le θ-quantile de régression pour $\boldsymbol{x}^\star$ s'écrit :

$$Q(\theta/\boldsymbol{x}^\star) = \sum_{i=1}^{n} \left(\boldsymbol{\lambda}_i^- - \boldsymbol{\lambda}_i^+ \right) K(\boldsymbol{x}_i, \boldsymbol{x}^\star) \qquad \text{et} \qquad \boldsymbol{w}_\theta = \sum_{i=1}^{n} \left(\boldsymbol{\lambda}_i^- - \boldsymbol{\lambda}_i^+ \right) \phi(\boldsymbol{x}_i), \tag{3}$$

où λ_i^-, λ_i^+ sont les multiplicateurs de Lagrange et $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ désigne une fonction noyau. Nous considérons ici la fonction noyau de type radial gaussien (RBF) donnée par :

$$k \left(\boldsymbol{x}_i, \boldsymbol{x}_j \right) = \exp \left(- \frac{|| \, \boldsymbol{x}_i - \boldsymbol{x}_j \, ||^2}{2 \, \sigma^2} \right), \tag{4}$$

où le paramètre σ désigne la taille de la fenêtre. Le paramètre σ peut être déterminé par validation croisée.

2.2 Détection de nouveauté par SVM

On peut étendre les méthodes de type SVM pour résoudre les problèmes de détection de nouveauté pour résoudre les problèmes de valeurs atypiques. L'approche de la détection de nouveauté en SVM (ou one-class SVM, Schölkopf et al. (1999)) consiste à créer une sphère de décision autour des données dans un espace transformé obtenu par la transformation ϕ. La méthode de détection de nouveauté en SVM va chercher à estimer cette sphère de décision en l'estimant à l'aide de support de vecteur. Le problème d'optimisation primale que l'on doit alors minimiser pour la détection de nouveauté est

$$\frac{1}{2} \|\boldsymbol{w}\|^2 - \rho + \frac{1}{n\nu} \sum_{i=1}^{n} \xi_i$$

sous les contraintes

$$< \Phi(x), \boldsymbol{w} > + b \geq \rho - \xi_i \quad et \quad \xi_i \geq 0, \quad i = 1, \ldots, n.$$

Le paramètre ν est utilisé pour contrôler le volume de la sphère et donc sert à contrôler le nombre de valeur atypique trouvé.

3 Application en chémoinformatique

Dans une étude précédente Villain et al. (2014a), nous avons établi des modèles basés sur 401 molécules. À partir de ces résultats, la solubilité moléculaire, le log(P) et la polarisabilité sont les 3 descripteurs fondamentaux associés au mode d'action des molécules. À partir de ces descripteurs, nous avons construit un modèle permettant de discriminer des molécules atypiques par le biais de modèle SVM de détection de nouveauté (voir Figure 1) afin de prédire le niveau de 36 médicaments testés au CERMN.

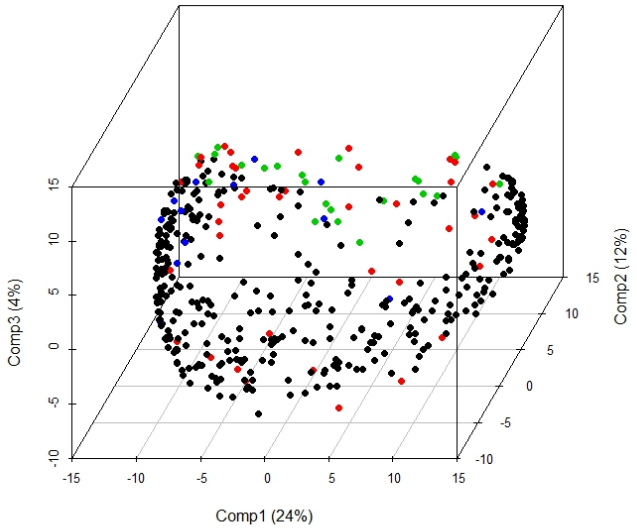

FIGURE 1 – Projection des 437 molécules sur les trois premières composantes principales de l'analyse en composante principale kernelisée.

La Figure 1 représente le positionnement des molécules sur les trois premières composantes principales d'une analyse principale kernelisée. Les points noirs et rouges représentent 401 molécules décrites dans Villain et al (2014a). Le points bleus et verts sont associés aux 36 médicaments dont 12 sont dans le domaine de validité (points bleus). Un intérêt de cette analyse est de pouvoir identifier les molécules se trouvant dans un espace de faible densité pour lesquelles nous avons peu d'information (points rouges et verts). On va maintenant s'intéresser dans la Figure 2 à la représentation graphique du biais des prédictions pour les 36 médicaments.

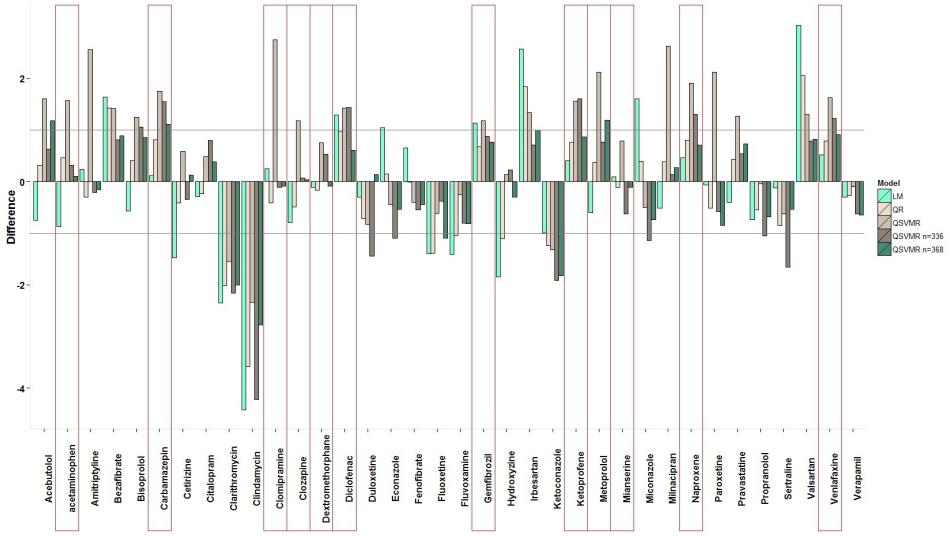

FIGURE 2 – Biais des prédictions du niveau de toxicité des médicaments.

La Figure 2 repésente le biais de prédiction des 36 médicaments pour les modèles de régression décrits dans Villain et al. (2014a). Les médicaments encadrés en rouge correspondent aux médicaments prédits comme étant dans le domaine de validité par notre modèle. Nous pouvons remarquer que la plupart des médicaments conservés ont un biais inférieur à une unité par rapport à la valeur réelle qui correspond à une erreur inférieur à 1 mol/L. De plus, nous observons une surestimation des prédictions qui est en écotoxicologie moins problématique qu'une sous estimation du niveau de toxicité.

	QSVM	QSVM $n = 336$	QSVM $n = 368$
EQM	2.05	1.48	0.86
EQM (avec domaine de validité)	1.3	0.87	0.69

En regardant le tableau ci-dessus, on peut souligner que l'utilisation d'un domaine de validé permet de diminuer l'erreur quadratique moyenne des modèles SVM et donc que l'utilisation d'une telle approche est intéressante.

Bibliographie

[1] Neuwoehner, J., Fenner, K., Escher, B. I. (2009), Physiologiccal Modes of Action of Fluoxetine and its human Metabolites in Algae, *Environmental Science & Technology*, 43, 6830-6837

[2] Schölkopf, B., Williamson, R.C., Smola, A.J., , Shawe-Taylor, J. and Platt, J.C. (1999) Support Vector Method for Novelty Detection. *Neural Information Processing System*, 12, 582-588.

[3] Vapnik, V.N. (1998), Statistical Learning Theory, New York.
[4] Villain, J. , Lozano, S., Halm-Lemeille, M.P., Durrieu, G. and Bureau, R. (2014a), Quantile regression model for a diverse st of chemicals : application to acute toxicity for green algae, *Journal of Molecular Modeling*, in press.
[5] Villain, J., Durrieu, G. and Bureau, R. (2014b),Quantile de réegression : application à l'analyse de l'écotoxicité de molécules chimiques, *46èmes journées de stastistique de la société française de statistique.*
[6] Villain, J., Minguez, L., Halm-Lemeille, M.P., Durrieu, G. and Bureau, R. (2015), Quantile regression models associated to acute toxicity for algae. Application for pharmaceutical compounds and characterization of their potential MOA., *Soumis.*