Prédiction de solubilité de molécules à partir des seules données relationnelles

Sébastien Derivaux, Agnès Braud, Nicolas Lachiche

LSIIT, ULP/CNRS UMR 7005
Pôle API, Bd Sébastien Brant - 67412 Illkirch, France
{derivaux,braud,lachiche}@lsiit.u-strasbg.fr

Résumé. La recherche de médicaments passe par la synthèse de molécules candidates dont l'efficacité est ensuite testée. Ce processus peut être accéléré en identifiant les molécules non solubles, car celles-ci ne peuvent entrer dans la composition d'un médicament et ne devraient donc pas être étudiées. Des techniques ont été développées pour induire un modèle de prédiction de l'indice de solubilité, utilisant principalement des réseaux de neurones ou des régressions linéaires multiples. La plupart des travaux actuels visent à enrichir les données de caractéristiques supplémentaires sur les molécules. Dans cet article, nous étudions l'intérêt de la construction automatique d'attributs basée sur la structure intrinsèquement multi-relationnelle des données. Les attributs obtenus sont utilisés dans un algorithme d'arbre de modèles, auquel on associe une méthode de *bagging*. Les tests réalisés montrent que ces méthodes donnent des résultats comparables aux meilleures méthodes du domaine qui travaillent sur des attributs construits par les experts.

1 Introduction

Pour créer un nouveau médicament, la pharmacologie opère en deux temps. Tout d'abord elle synthétise un grand nombre de molécules. Ces molécules sont ensuite appliquées sur un substrat simulant la pathologie que le médicament recherché doit combattre. Le débit de molécules synthétisées puis testées a grandement augmenté ces dernières décénnies avec l'introduction de la synthèse combinatoire et le criblage à haut débit (Hou et al., 2004). Ce processus peut néanmoins être encore amélioré. En effet, une propriété essentielle des médicaments est de pouvoir être solubles pour circuler à travers le système sanguin afin d'atteindre la partie malade de l'organisme, or cette propriété n'est pas vérifiée par toutes les molécules. Idéalement, les molécules non solubles ne devraient être ni testées ni même synthétisées afin d'accélérer le processus.

La solubilité d'une molécule est représentée par un attribut numérique nommé indice de solubilité. Les laboratoires pharmacologiques connaissent cette valeur pour un grand nombre de molécules. Ceci motive l'utilisation de méthodes issues de la fouille de données pour induire un modèle qui, à partir de la structure d'une molécule, prédit son indice de solubilité.

Dans le cadre de cette application, une base de données permet de décrire les molécules à partir de trois tables :

– La table molécule contient les caractéristiques globales de la molécule, réduites, dans notre cas, à un identifiant de la molécule et à l'indice de solubilité. Ce dernier est l'attribut cible de notre tâche de régression. Chaque n-uplet de cette table correspond donc à un exemple.
– La table atome contient la description de tous les atomes composant les molécules. Chaque atome est ainsi décrit par l'identifiant de la molécule à laquelle il appartient, un identifiant d'atome et son symbole atomique (carbone, azote, . . .).
– La table liaison contient la description des liaisons des molécules. Chacun de ses n-uplets est défini par les deux identifiants des atomes liés entre eux et le type de liaison (simple, double, . . .).

Initialement, la table molécule ne comprend aucun attribut numérique ou nominal autre que l'attribut cible lui-même. Les seuls attributs non-structurels sont le type des atomes et le type des liaisons. Il est nécessaire d'enrichir les données car pour donner de bons résultats, les méthodes de régression ont besoin d'attributs numériques. La méthode classique est d'ajouter des attributs que les experts savent être en corrélation avec la solubilité. Le problème de cette méthode est qu'elle demande des connaissances pointues dans le domaine de la chimie moléculaire, et nécessite d'investir dans des logiciels coûteux, capables d'obtenir ces descripteurs comme le montre l'état de l'art de Delaney (2005). De plus certains de ces descripteurs, comme le *log P* représentant la lipophilie d'une molécule ne sont que des prédictions. C'est pour ces raisons que nous proposons une méthode de construction automatique d'attributs, notamment numériques.

Dans cet article, nous commençons par proposer une méthode de construction automatique d'attributs. Nous détaillerons ensuite notre algorithme de modèle d'arbre couplé à du *bagging* qui, à partir des attributs précédemment créés, permet d'induire un modèle de régression. Enfin, nous étudions les performances de notre approche, en démontrant l'apport de chacune des techniques utilisées et en comparant nos modèles induits avec les meilleurs de la littérature sur ce sujet.

2 Construction automatique d'attributs

L'enrichissement manuel des données demandant du temps, des compétences et des logiciels couteux, nous proposons donc une technique pour construire automatiquement des attributs à partir de la structure des molécules.

Jusqu'à récemment, la construction des nouveaux attributs à partir de données relationnelles se basait soit sur la sélection (Kramer et al., 2001) qui construit des attributs booléens (par exemple *est-ce que la molécule a une liaison carbone-oxygène ?*), soit sur l'agrégation (Perlich et Provost, 2003) qui contruit des données nominales ou numériques (par exemple *le nombre d'atomes de la molécule*). Seule l'union de l'agrégation et de la sélection permet d'exprimer des attributs comme : *nombre de liaisons carbone-oxygene*.

Vens et al. (2003) sont les seuls, à notre connaissance, qui combinent l'aggrégation et la sélection. Ils utilisent des forêts aléatoires pour effectuer une sélection dans l'espace des attributs constructibles.

Nous nous proposons d'utiliser les graphes de sélections afin de définir des attributs constructibles. Les graphes de sélection introduits par (Knobbe et al., 1999) permettent d'exprimer graphiquement des motifs sur des données multi-relationnelles. Un motif est un ensemble de

caractéristiques de structure qui peuvent être, ou non, vérifiées par une molécule. Un exemple est donné à la figure 1 représentant une liaison (de type quelconque) entre un atome de carbone et un atome d'oxygène.

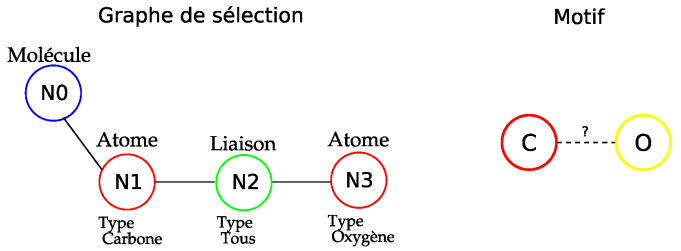

FIG. 1 – *Exemple de graphe de sélection à gauche avec son interprétation comme motif chimique à droite.*

Avant de créer les nouveaux attributs, nous cherchons les motifs dans les données. Pour ce faire, nous construisons les graphes de sélection suivant une grammaire simple. Les graphes de sélection sont composés d'un noeud molécule, suivi d'une alternance de noeuds atome et liaison. Au niveau des motifs cela permet d'exprimer des séquences d'atomes. Pour des questions de calculabilité, nous nous limitons à des séquences d'au plus 2 atomes. En effet, nos tests ont montré que des séquences plus longues (jusqu'à 4 atomes) n'améliorent pas le résultat.

L'algorithme considère l'ensemble des graphes de sélection constructibles pour les données, et pour chacun construit un nouvel attribut numérique. La valeur des attributs construits sera, pour chaque exemple, le nombre d'occurrences du graphe de sélection.

L'algorithme construit également des attributs plus complexes. Pour ce faire, nous procédons comme précédemment, mais au lieu de se placer au niveau de la molécule, nous nous plaçons au niveau des atomes et des liaisons en ajoutant, par exemple, pour chaque atome, un attribut représentant le nombre de ses liaisons avec d'autres atomes. Il faut néanmoins noter que ces attributs ne sont pas des descriptions en soit de la molécule mais uniquement des composants de la molécule. Pour répercuter l'information au niveau de la molécule, nous utilisons les opérateurs classiques d'agrégation suivants : moyenne, somme, minimum, maximum. Ainsi, on peut ajouter à la description de la molécule le nombre moyen de liaisons de chacun de ses atomes par exemple.

3 Algorithme d'induction de modèles de régression

Le problème de prédiction de l'indice de solubilité est généralement résolu en utilisant soit un réseau de neurones multicouches (Tetko et al., 2001; Huuskonen, 2000) soit une régression linéaire multiple (Hou et al., 2004; Delaney, 2004).

Dans le domaine de la fouille de données, d'autres algorithmes ont été développés et donnent de bons résultats, nous nous intéressons particulièrement aux algorithmes à base d'arbres.

Mr-SMOTI (Appice et al., 2003) permet de créer un arbre de décision où les feuilles sont constituées de régression linéaires multiples. De plus, Mr-SMOTI a la capacité de travailler sur des données multi-relationnelles. Néanmoins, il se limite à utiliser les attributs numériques de la table cible pour induire ses régressions linéaires multiples et ne construit par sélection que des attributs booléens pour les noeuds de partitionnement.

Dans l'approche proposée, l'algorithme induit un arbre de modèles où les feuilles et les noeuds internes jouent des rôles complémentaires. Une feuille effectue la tâche de régression proprement dite. A partir d'un ensemble d'exemples, la feuille effectue une régression linéaire multiple avec les attributs autorisés par ses noeuds ascendants. Une feuille racine n'a accès à aucun attribut. Un noeud interne, appelé un raffinement, a pour but de simplifier la tâche de sa descendance. Pour cela, un noeud peut soit partitionner l'espace des exemples (cas typique des arbres de décision) selon un seuil sur un des attributs numériques construits, soit proposer un nouvel attribut qui sera accessible à sa descendance. Le nombre de fils d'un noeud dépend du type de raffinement effectué, un raffinement partitionnant l'espace des exemples selon une condition booléenne aura deux fils, tandis qu'un noeud introduisant un nouvel attribut n'aura qu'un fils. La construction de l'arbre se déroule de la façon suivante. L'algorithme débute en créant une feuille initiale. Comme elle n'a accès initialement à aucune variable numérique, la régression linéaire se réduit à une constante. Ensuite cette feuille est raffinée, l'algorithme teste tous les raffinements possibles, c'est-à-dire tous les ajouts d'attributs et tous les partitionnement possibles. Enfin le raffinement améliorant au mieux l'erreur du modèle est conservé.

L'arbre ainsi induit est bien souvent sujet au sur-apprentissage. Dans le cas des arbres, une façon de palier le sur-apprentissage est de procéder à un élagage de l'arbre obtenu. Nous utilisons l'élagage par réduction d'erreur définie par Quinlan (1987) avec un jeu de validation représentant un tiers du jeu d'apprentissage.

Enfin nous utilisons la technique du *bagging* (Breiman, 1996) en induisant 50 modèles avec chacun une partition jeu d'apprentissage et jeu de validation différent, ce qui donne une indépendance aux modèles induits. La prédiction finale est la moyenne des prédictions des 50 modèles.

4 Résultats

Les jeux de données utilisées dans nos tests correspondent à des ensembles de molécules de différents types dont la solubilité est connue. Ces données nous ont été fournies par le laboratoire d'Infochimie ULP/CNRS UMR 7551. Dans tous les cas les résultats sont obtenus à l'issue d'une validation croisée en 10 partitions.

Les premiers test ont été réalisés sur un jeu de 511 molécules. Les résultats obtenus sont présentés dans la table 1. On peut remarquer que l'utilisation conjointe du *bagging* et des arbres de modèles permet de réduire sensiblement l'erreur. Le *bagging* donne de meilleurs résultats sur les arbres de modèles que sur les régressions linéaires multiples.

Nous avons également comparé nos résultats à ceux obtenus dans le milieu de l'infochimie. Pour ce faire, nous avons utilisé un jeu d'apprentissage de 1635 molécules et le jeu de Yalokowsky (Yalkowsky et Banerjee, 1991) comme jeu de test. Le jeu de test comprend 21 molécules définies comme représentatives par Yalkowsky et est communément utilisé comme jeu de comparaison. Nous avons reporté dans la table 2 les résultats de quatre méthodes parmi les plus performantes. Ces méthodes utilisent différents attributs pour décrire les molécules

Méthode	Erreur moyenne	Racine de l'erreur quadratique
Régression linéaire	0.66	0.92
Régression linéaire + Bagging	0.64	0.88
Arbre de modèles	0.58	0.86
Arbre de modèles + Bagging	0.52	0.78

TAB. 1 – *Résultats obtenus en utilisant différentes configurations.*

dont notamment des attributs d'états électroniques, d'états d'hybridations et l'indice de lipophilie. Certains de ces attributs nécessitent des logiciels commerciaux spécialisés. La précision de notre modèle induit est assez proche de celles des modèles induits en utilisant des connaissances expertes en plus.

Référence	Méthode	Racine de l'erreur quadratique
Huuskonen (2000)	Réseau de neurones	0.63
Tetko et al. (2001)	Réseau de neurones	0.64
Hou et al. (2004)	Régression linéaire multiple	0.64
Delaney (2004)	Régression linéaire multiple	0.78
Notre approche	Arbre de modèles + Bagging	0.80

TAB. 2 – *Résultats de notre approche sur le jeu de test de Yalkowsky*

5 Conclusion

Dans cet article nous avons proposé une nouvelle méthode d'induction de modèles de prédiction de la solubilité des molécules. Cette méthode se base sur l'utilisation de techniques nouvelles dans ce domaine d'application, comme les arbres de modèles et le *bagging*. Nous proposons également une méthode pour utiliser les données multi-relationnelles brutes, sans l'ajout d'attributs experts, par construction automatique d'attributs numériques en utilisant des opérateurs d'agrégation. Nos résultats sont proches de ceux obtenus par les meilleurs approches développées jusqu'à présent, approches utilisant des attributs experts.

Plusieurs voies d'évolution sont possibles. La première consisterait à augmenter la capacité de construction d'attributs en ne la limitant plus à des motifs simples. Cela peut passer par l'utilisation d'éléments de plus haut niveau dans les motifs, comme l'utilisation des cycles aromatiques (plusieurs atomes de carbone formant un cercle). Une seconde voie consiste en l'extension des arbres de modèles qui, plutôt que de se limiter à des régressions linéaires multiples, utiliseraient, quand c'est indiqué, des réseaux de neurones de topologie simple.

Références

Appice, A., M. Ceci, et D. Malerba (2003). Mining model trees : A multi-relational approach. In *ILP*, Volume 2835 of *Lecture Notes in Computer Science*, pp. 4–21.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*(2), 123–140.

Delaney, J. S. (2004). ESOL : Estimating aqueous solubility directly from molecular structure. *Journal of Chemical Information and Computer Sciences 44*(1), 1000–1005.

Delaney, J. S. (2005). Predicting aqueous solubility from structure. *Drug Discovery Today 10*, 289–295.

Hou, T., K. Xia, W. Zhang, et X. Xu (2004). Adme evaluation in drug discovery. 4. prediction of aqueous solubility based on atom contribution approach. *Journal of Chemical Information and Computer Sciences 44*(1), 266–275.

Huuskonen, J. (2000). Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *Journal of Chemical Information and Computer Sciences 40*(3), 773–777.

Knobbe, A. J., H. Blockeel, A. P. J. M. Siebes, et D. M. G. van der Wallen (1999). Multi-relational data mining. Technical Report INS-R9908.

Kramer, S., N. Lavrac, et P. Flach (2001). Propositionalization approaches to relational data mining. In S. Dzeroski et N. Lavrac (Eds.), *Relational Data Mining*, pp. 262–291. Springer-Verlag.

Perlich, C. et F. Provost (2003). Aggregation-based Feature Invention and Relational Concept Classes. *Proceedings of the Ninth International Conference on Knowledge Discovery and Data Mining (KDD-2003)*, 167–176.

Quinlan, J. R. (1987). Simplifying decision trees. *Int. J. Man-Mach. Stud. 27*(3), 221–234.

Tetko, I. V., V. Y. Tanchuk, T. N. Kasheva, et A. E. P. Villa (2001). Estimation of aqueous solubility of chemical compounds using e-state indices. *Journal of Chemical Information and Computer Sciences 41*(6), 1488–1493.

Vens, C., A. Van Assche, H. Blockeel, et S. Dzeroski (2003). First Order Random Forest with Complex Aggregates. In *ILP*, Volume 2835 of *Lecture Notes in Computer Science*, pp. 323–340.

Yalkowsky, S. H. et S. Banerjee (1991). *Aqueous Solubility : Methods of Estimation for Organic Compounds*. Marcel Dekker.

Summary

The search for new drugs passes by the synthesis of candidate molecules whose effectiveness is then tested. This process can be speeded up by identifying the nonsoluble molecules, because those cannot enter into the composition of a drug and thus should not be tested. Techniques were developed, which mainly use neural networks or multiple linear regressions, in order to induce a predictive model of the aqueous solubility. Most current works aim at enriching the data with additional characteristics on the molecules. In this article, we rather study the interest of the automatic construction of attributes based on the intrinsically multi-relational structure of the data. The attributes obtained are used in an algorithm for model tree induction, which is associated with a bagging method. The tests carried out show that our method gives results comparable with the best methods of the field, which work on attributes built by experts.