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 :