Micro-projet : prédiction de médicaments

ES Apprentissage Artificiel

2016

1 Données

Les médicaments agissent dans la plupart des cas à travers une ou des interactions avec une ou des protéines, dites cibles, intervenant dans une ou des voies métaboliques nécessaires pour le développement de la maladie. Une fois qu'une cible a été identifiée pour une maladie, le processus menant à la découverte d'un médicament interagissant avec la cible est complexe, long et coûteux ne serait-ce que pour l'optimisation, la synthèse et la certification de la molécule.

Pour réduire ces coûts, on peut s'aider de la prédiction de l'intéraction de médicaments avec de futures cibles thérapeutiques. On propose dans ce projet de prédire l'intéraction de médicament avec des protéines appartenant à diverse familles protéiques.

On utilise le graphe moléculaire pour calculer un noyau de similarité entre molécules sur la base de similarité entre leur deux graphes. Concernant les protéines, on utilise un noyau basée sur la similarité de leur séquence d'acides aminées. Le 'kernel trick' permet dans ce projet de travailler des objets chimiques et biologique.

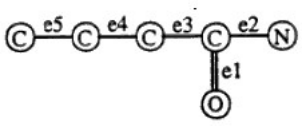

Figure 1: graphe moléculaire du glucagon

Glucagon:
His−Ser−Gin−Gly−Thr−Phe−Thr−Ser−
Asp−Tyr−Ser−Lys−Tyr−Leu−Asp−Ser−
Arg−Ala−Gin−Asp−Phe−Val−Gin−Trp−
Leu−Met−Asn−Thr−

Figure 2: séquence d'acides aminées du glucagon

Le jeu de donnée contient les fichiers suivants:

- noyau de similarité des molécules basé sur la similarité entre leurs graphes
- noyau de similarité des protéines basé sur la similarité entre leurs séquences d'acides aminées
- fichier tsv détaillant le profile de ligands pour chaques protéines. Les valeurs du tableau correspondent à 1 si la protéine et la molécule (correspondant respectivement à la ligne et à la colonne) intéragissent, et O sinon.

2 Problème

2.1 Support Vector Classifier

1. Pour chaque molécule, séparer son jeu de données en deux jeux de taille égale.
2. Pour chaque molécule, en utilisant la mesure de similarité basée sur les séquences des protéines:
   (a) Entrainer un SVC sur le premier jeu de données et prédire l'intéraction entre chaque molécule avec les protéines du second jeu de données.
   (b) Qn faisant varier la constante C du SVC, évaluer la performance du SVC en termes d'exactitude (accuracy).
3. Pourquoi ne pas entrainer puis évaluer le SVM sur tout le jeu de données ? Quelle autre procédure de division de jeu de données pourrait-on utiliser pour valider la performance du SVC ?
4. Quelles autres mesures pourrait-on utiliser pour évaluer la performance du SVC ? Laquelle ou lesquelles vous paraissent pertinentes pour évaluer la prédiction de l'intéraction entre des médicaments et des cibles protéiques ?
5. Au lieu d'utiliser la similarité entre protéine sur la base de leur séquence, on propose maintenant d'étudier la performance du SVC en considérant la similarité du profile de ces protéines. Le profile d'une protéine est un vecteur composé de 1 ou 0 à chaque indice j selon que la protéine intéragisse ou non avec la j-ème molécule du dataset. Les protéines étant maintenant représentées par un vecteur, on peut appliquer un SVC avec noyau RBF gaussien (le coefficient σ^2 pertinent étant le nombre moyen de ligand par protéine) pour prédire l'intéraction des molécules avec les protéines.
   (a) Pourquoi cette représentation des protéines est-elle pertinente?
   (b) Répéter la procédure de la question 2 en utilisant le profile des protéines comme représentation de celles-ci avec un SVC-noyau RBF gaussien.
   (c) Comparer les performances du SVM selon les 2 mesures de similarités testées. Quelle erreur a été commise et comment aurait-il fallu procéder pour comparer les deux représentations protéiques de façon équitable ?

2.2 Régression Linéraire Régularisé pour la classification

Les méthodes de régression peuvent être utilisées dans un contexte de classification. Concernant la classification à 2 classes, assigner la première classe à la valeur réelle 1 et la seconde à la valeur réelle 0 et considérer un une valeur seuil entre 0 et 1 séparant les 2 classes permet d'appliquer directement une méthode de régression à un problème de classification.

1. Répéter la procédure de la question 2 en utilisant la méthode de régression linéaire kernelisé avec régulairsaton L2 ("Kernel Ridge Regression"). Comparer les performances en terme de score "ROC-AUC".
2. Quelle est l'intérêt de la régularisation L2 ?
3. Quels sont les avantages et les inconvénients d'utiliser cette méthode plutôt qu'un SVC pour un problème de classification ?

2.3 Optionnel : apprentissage multi-tache

Dans les sections 2.1 et 2.2, nous avons prédits les cibles protéiques des molécules une par une par des méthodes d'intelligence artificielle en faisant l'hypothèse du paradigme suivant : si une première protéine a une séquence d'acide aminé 'fortement' similaire à celle d'une seconde protéine ciblée par la molécule, alors il y a de 'forte' chance que la molécule cible aussi la première protéine.
On peut faire la même hypothèse du point de vue des protéines : une molécule ayant un graphe moléculaire 'fortement' similaire à celui d'une molécule ciblant une protéine a de 'forte' chance de cibler cette même protéine.

Une approche plus sophistiquée serait ainsi d'utiliser la similarité entre molécule combinée avec la similarité entre protéine pour prédire l'intéraction entre protéines et molécules [1]. Cette approche est appelé "approche multi-tache" : au lieu de prédire l'intération de protéines avec une molécule en utilisant uniquement l'information disponible pour cette molécule (ce qui correspond à une tache), on utilise les informations disponibles de plusieurs molécules pour prédire les cibles protéiques d'une molécule en relayant ces informations à travers la valeur de similarité entre molécule.
Suivant [2], l'approche multi-tache revient à utiliser la méthode à noyau en considérant le kernel résulant du produit de kronecker du kernel de molécule et du kernel de protéine.

1. Calculer le produit de kronecker des 2 kernels.
2. En concaténant les séparations en 2 datasets effectué pour chaque molécule lors de la question 2.1.1, utiliser le premier dataset généré pour entrainer un SVC et prédire sur le second dataset généré.
3. Comparer les performances de cette approche par rapport à celle de la question 2.1.2.b. Commenter.

References

[1] Twan van Laarhoven, Sander B Nabuurs, and Elena Marchiori. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics*, 27(21):3036–3043, 2011.

[2] Laurent Jacob and Jean-Philippe Vert. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics*, 24(19):2149–2156, 2008.