Intelligente Datenanalyse

Prüfung: Projektaufgabe 10

Paul Prasse
Dr. Niels Landwehr
Prof. Tobias Scheffer

Die Projektaufgabe ist Teil der Prüfung *Intelligente Datenanalyse in Matlab*. Jede Aufgabe soll durch einen Studenten selbstständig bearbeitet und die Lösung innerhalb der mündlichen Prüfung vorgestellt werden. Ein Ausdruck des MATLAB-Programmcodes und der Ergebnisse in Form eines Diagramms, Tabelle o.ä. werden vorausgesetzt; die Art der Präsentation der Ergebnisse ist dem Studenten freigestellt.

Problemstellung

Ein Pharmaunternehmen möchte ein neues Medikament entwickeln welches die Wirkung eines bestimmten Hormons unterbindet. Damit ein Hormon in der Zelle wirksam werden kann, muss es sich an einen hormonspezifischen Rezeptor der Zelle binden. Ziel ist es daher diese Rezeptoren zu blockieren. Dies geschieht mit so genannten *Liganden* – kleine Moleküle welche sich an einen Rezeptor binden. Um zu prüfen ob ein bestimmtes Molekül als Ligand geeignet ist werden aufwendige Labor-Experimente durchgeführt. Dabei wird die Bindungsfähigkeit bestimmt; je höher diese ist desto geeigneter ist das Molekül um als Teil eines neuen Wirkstoffes zu dienen. Allerdings gibt es nahezu unendlich viele Moleküle welche als Ligand in Frage kommen. Um Zeit und Geld für die Entwicklung zu sparen, ist das Unternehmen an einer möglichst genauen Vorhersage der Bindungsfähigkeit interessiert. Basierend auf der Vorhersage sollen dann eine kleine Zahl von Kandidaten ausgewählt und nur für diese Moleküle Labor-Experimente durchgeführt werden.

Für 232 Moleküle wurde die Bindungsfähigkeit – der natürliche Logarithmus der *Relative Binding Affinity* (logRBA) – experimentell bestimmt. Jedes dieser Moleküle wird durch einen SMILES-Code repräsentiert und kann bspw. auf der Webseite **http://pubchem.ncbi.nlm.nih.gov/edit** visualisiert werden. Zusätzlich zum SMILES-Code und dem logRBA-Wert ist die Struktur eines Moleküls wie folgt gegeben: Jedes Molekül wird als ungerichteter Graph interpretiert. Die Knoten des Graphen sind die Atome des Moleküls. Besteht zwischen zwei Atomen eine Einfachbindung entspricht dies einer Kanten mit dem Gewicht 1; besteht eine Doppelbindung wird die zugehörige Kante mit 2 gewichtet; andernfalls ist keine Kante zwischen den betreffenden beiden Knoten. Abbildung 1 zeigt beispielhaft die Informationen für ein Molekül. Sie wurden damit beauftragt ein Vorhersagemodell zu entwickeln welches für ein neues Molekül, basierend auf dessen Struktur, die Bindungsfähigkeit prognostiziert.

Aufgabe

Importieren sie die Daten in MATLAB, wählen sie eine geeignete Datenrepräsentation und führen sie falls nötig eine Datenvorverarbeitung durch. Diskutieren und implementieren sie ein Ähnlichkeitsmaß welches die Strukturähnlichkeit zweier Moleküle bewertet. Überlegen sie welche Art von Analyseproblem vorliegt. Identifizieren und implementieren sie ein geeignetes Verfahren zum Lösen des Problems unter Verwendung des gewählten Ähnlichkeitsmaßes. Trainieren und evaluieren sie ein Modell basierend auf den gegebenen

```
SMILES: CC1CCCC(O)CCCCCc2cc(O)cc(O)c2C(=O)O1
logRBA: 1.480000000000000e+000
  2   1   1
  3   2   2
  4   2   1
  5   1   1
  6   3   1
  7   1   2 <-- Zweifachbindung zwischen dem 7. und 1. Atom
  8   4   2
  9   8   1
 10   3   1
 11   5   1 <-- Einfachbindung zwischen dem 11. und 5. Atom
 12   9   1
 13   4   1
 14  15   1
 15  18   1
 16  17   1
 17  11   1
 18  16   1
 19  11   1
 20  15   1
 21  13   1
 22  23   1
 23  21   1
  9   6   2
 20  22   1
```

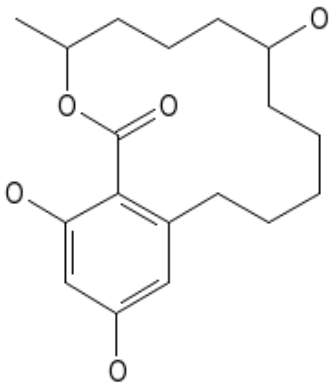

Abbildung 1: Beispiel-Informationen über ein Molekül.

Daten (molecule.txt). Begründen und dokumentieren sie kurz alle durchgeführten Schritte.