Einfluss der Gewichtung von Moleküldeskriptoren auf die Klassifizierung von Arzneistoffen

S.F.Badreddin Abolmaali, Andreas Zell

Eberhard-Karls-Universität Tübingen, Wilhelm-Schickard-Institut für Informatik
Köstlinstr. 6, D-72074 Tübingen
{abolmaali, zell}@informatik.uni-tuebingen.de

Moleküldeskriptoren dienen dazu, chemische Verbindungen problemspezifisch zu charakterisieren. Klassifizierungsverfahren werden in der pharmazeutischen Industrie u.a. dazu verwendet, Moleküldatenbanken nach potentiellen Arzneistoffen bzw. Leitstrukturen zu durchsuchen bzw. Arzneimittel verschiedener Indikationsgruppen voneinander abzugrenzen. In Bezug auf die Diskriminierung unterschiedlicher Verbindungsklassen verhalten sich die einzelnen Moleküleigenschaften und damit die ihnen zugeordneten Deskriptoren allerdings nicht homogen. So sind Deskriptoren, die bei Verbindungen unterschiedlicher Klassen ähnliche Werte annehmen, für die Diskriminierung weniger aussagekräftig als solche, die für die meisten Verbindungen eindeutige Werte liefern. Inwiefern eine differenzierte Gewichtung der Deskriptoren entsprechend ihrer Aussagekraft eine verbesserte Unterscheidung verschiedener Verbindungsklassen ermöglicht, wurde im Rahmen des BMBF-Projektes SOL (Suche und Optimierung von Leitstrukturen) [SOL99] untersucht. Hierzu wurden zwei verschiedene Moleküldatensätze mit dem Klassifizierungsverfahren LVQ (Learning Vector Quantization) sowie einem zweistufigen feedforward Netzwerk (Multi-Layer-Perzeptron, MLP) diskriminiert.
Die verwendeten Moleküldatensätze umfassen 5033 bzw. 11590 Einträge. In beiden sind Arzneistoffe und nicht-Arzneistoffe je zu ca. 50% vertreten. Als Deskriptoren wurden in einem Fall die Anzahl von Substrukturen (28 Deskriptoren), im anderen Fall physikochemische Eigenschaften (195 Deskriptoren) verwendet. Die Datensätze wurden je in einen Trainings- und zwei Testdatensätze im Verhältnis 90:5:5 bzw. 80:10:10 aufgeteilt.
Im Fall von LVQ [Kohonen95] wird zunächst die Dichteverteilung der Eingabevektoren des Trainingsdatensatzes mit Hilfe von sogenannten Code-Book Vektoren approximiert. Die Dimensionalität der Vektoren entspricht der Anzahl der verwendeten Deskriptoren. In einem zweiten Schritt wird sukzessive jeder Molekülvektor des Testdatensatzes einem trainierten Code-Book Vektor und damit einer bestimmten Klasse zugeordnet.
Das MLP-Netzwerk der Topologie 28-6-1 wurde im SNNS (Stuttgart Neuronal Network Simulator) [Zell99] realisiert. Es wurde in 200 Lernzyklen mit der Lernfunktion Scaled Conjugate Gradient (SCG) trainiert.
Die Untersuchungen mit LVQ ergaben bei den beiden Datensätzen Klassifizierungsraten von 78 bzw. 67% bei nicht gewichteten Deskriptoren. Bei einer Gewichtung in Grössenordnungen von 1-2 bis 1-100.000 ergaben sich Werte zwischen 70 und 84% bzw. 65 und 70%. Diese Ergebnisse beruhen auf unterschiedlichen Gewichtungsmustern, sind jedoch unabhängig vom Wertebereich der Gewichtungsfaktoren. Zur Klärung dieses Effekts wurden die initialen Code-Book Vektoren ohne vorherige Gewichtung der Deskriptoren verändert. Hierbei ergaben sich Klassifizierungsergebnisse von 72 bis 81% bzw. 65 bis 69%, was darauf hindeutet, dass die Gewichtung der Deskriptoren eine Verschiebung der initialen Code-Book Vektoren – relativ zu den übrigen Vektoren des Eingaberaumes – darstellt. Bei der Skalierung zweier verschiedener Testdatensätze mit den gleichen Gewichtungsmustern ergab sich eine schwache Korrelation der Klassifizierungsergebnisse. Das Quadrat des Pearsonschen Korrelationskoeffizienten betrug im einen Fall 0,14, im anderen lediglich 0,0005. Demnach ist es nicht möglich, ein optimales Gewichtungsmuster für unterschiedliche Testdatensätze zu ermitteln, und so die Diskriminierung verschiedener Moleküldatensätze zu verbessern. Diese Ergebnisse stimmen mit der mathematischen Aussage überein, nach der Klassifizierungsverfahren, die die Dichteverteilung der Eingaberaumes gleichmässig approximieren, von Skalierung unabhängig sind.
Die Klassifizierung des auf Substruktur-Counts basierenden Moleküldatensatzes mit dem oben beschriebenen MLP-Netzwerk ergab ein vollkommen anderes Bild. Die Klassifizierungsergebnisse sind hier sowohl vom jeweiligen Gewichtungsmuster als auch vom Wertebereich der Gewichtungsfaktoren abhängig. Ohne Gewichtung erreicht das Verfahren eine Klassifikation von 100%. Eine Gewichtung mit Faktoren von 1-100.000 verschlechtert das Ergebnis auf 59-100%. Die Klassifizierungsraten zweier verschiedener Testdatensätze, die mit den gleichen Gewichtungsmustern skaliert wurden, korrelieren hier – wiederum abhängig vom Wertebereich der Gewichtung (1-100 bzw. 1-100.000) – mit einem Faktor von 0,98 bzw. 0,55. Die Klassifikationsergebnisse liegen hier zwischen 59 und 100% bzw. 39 und 70%. Es hat sich gezeigt, dass die Diskriminierung umso eindeutiger ausfällt, je näher die Eingabedaten am Wendepunkt der Aktivierungsfunktion liegen. Diese Ergebnisse sprechen dafür, Deskriptordaten erst nach einer geeigneten Normierung für die Klassifikation mit MLP-Netzwerken zu verwenden.

[SOL99] http://www-ra.informatik.uni-tuebingen.de/forschung/sol/welcome.html
[Kohonen95] Kohonen et. al., LVQ_PAK, The Learning Vector Quantization Program Package, Version 3.1, User Manual (1995), Finnland
[Zell99] Zell et. al., SNNS, Stuttgart Neuronal Network Simulator, Version 4.2, User Manual (1999), Tübingen