Illustration: Fotolia / freshideas

Von lernfähigen Maschinen lernen

VON ANNA SACHER UND FABIAN THEIS, MÜNCHEN

Bioinformatiker bringen Computern bei, biologische Fragen zu lösen. Maschinelle Lernprogramme helfen den Rechnern dabei auf die Sprünge

Heute ist es selbstverständlich, dass Maschinen und Roboter den Menschen bei der täglichen Arbeit unterstützen. Wie den Menschen, muss man aber auch Maschinen die einzelnen Arbeitstechniken und Handgriffe erst einmal beibringen. Einen Fließbandroboter so zu programmieren, dass er das richtige Teil an der vorgesehenen Stelle einbaut, ist inzwischen keine große Herausforderung mehr. Kompliziert wird es, wenn eine Maschine lernfähig sein soll, um wie ein Mensch auf unbekannte Situationen intelligent zu reagieren. Hier helfen nur höhere Mathematik und komplexe Programmierung.

In den letzten Jahren hat dieses sogenannte maschinelle Lernen jedoch große Fortschritte gemacht. Ein Erfolgsbeispiel ist der Computer AlphaGo, der kürzlich einen Profi-Spieler des Strategiespiels Go besiegte – Experten hatten zuvor prognostiziert, dass es noch „ein weiteres Jahrzehnt“ dauern würde, bis eine intelligente Maschine einen Profi-Go-Spieler bezwingt. Das IBM-Computerprogramm Watson, das ebenfalls auf maschinellem Lernen basiert, übertrumpfte bereits 2011 in der Quizshow Jeopardy zwei Jeopardy-Champions. Inzwischen soll Watson sogar in der Medizin eingesetzt werden und Ärzten helfen, Behandlungsstrategien schneller und effizienter zu planen.

> „Kompliziert wird es, wenn eine Maschine lernfähig sein soll, um wie ein Mensch auf unbekannte Situationen intelligent zu reagieren.“

Auch wenn das Wort „Naturwissenschaften“ erst einmal den Einsatz künstlicher Ressourcen wie Computer und Maschinen ausschließt, sind diese insbesondere in der Biologie und Biomedizin enorm wichtig. Erst mit Hilfe künstlicher Intelligenz gelingt es, Erklärungen für die komplexen Vorgänge in der Biologie zu finden, die vom Menschen nur bis zu einem gewissen Grad verstanden und verarbeitet werden können. Gerade wenn Informationen aus vielen verschiedenen Richtungen eintreffen oder miteinander kombiniert und integriert werden müssen, reicht die Leistung des menschlichen Gehirns nicht aus. Um in diesen Fällen entsprechende Fragestellungen zu lösen benötigen wir die Unterstützung von Maschinen.

Hier zeigt sich auch, was Mathematiker, Statistiker und Informatiker in der Biologie ‚verloren’ haben: sie verknüpfen komplexe Vorgänge miteinander, erkennen Muster und leiten daraus Gesetzmäßigkeiten ab. Maschinelles Lernen ist ein essentieller Bestandteil der rechnerbasierten Biologie (Computational Biology). Dies umso mehr, seit im Zuge von Big Data enorme Datenmengen in der biologischen und biomedizinischen Forschung gesammelt werden, die mit klassischen Methoden der Datenverarbeitung meist nicht ausgewertet werden können.

Für die Entwicklung des maschinellen Lernens ist die große Datenmenge jedoch eher Segen als Fluch: Je mehr der Computer mit Informationen gefüttert wird, desto besser kann er lernen. Dem Menschen fällt es schwer, aus einem Berg von Daten essentielle Bestandteile heraus zu filtern und vor allem Zusammenhänge festzustellen. Der Computer erkennt dagegen Muster in den Daten, speichert sie, verfeinert die weitere Suche und entwickelt letztlich Gesetzmäßigkeiten. Maschinelles Lernen erfolgt über Algorithmen, die Schritt für Schritt während des Durchforstens der Muster erstellt werden. Über diese Algorithmen gelangt man schließlich zur Regel – und über diese zum Gesetz, das verifiziert oder widerlegt wird.

„Je mehr der Computer mit Informationen gefüttert wird, desto besser kann er lernen.“

Wichtige Methoden des maschinellen Lernens sind künstliche neuronale Netze, Regressions-und Klassifikationsmodelle, Random Forests, Support Vector Machines, Algorithmen zur Matrixfaktorisierung sowie graphische Modellierung. Man unterscheidet hierbei zwischen zwei hauptsächlichen Richtungen: überwachtes sowie unüberwachtes Lernen.

Ziel des überwachten Lernen ist es, einen funktionellen Zusammenhang zwischen Eingabe- und Zieldaten zu erkennen. Der Maschine wird dieser Zusammenhang anhand von Beispielen mittels Ein-/Ausgabepaaren beigebracht. Beim unüberwachten Lernen hingegen lernt der Computer ohne Zielvorgaben die Struktur eines Datensatzes und findet selbst Regeln und Muster.

Zum überwachten Lernen gehören zum Beispiel Klassifikationsmodelle, die unter anderem benutzt werden, um Daten in vorgegebene Kategorien einzuteilen. Ein weiteres Beispiel sind Regressionsanalysen, die unabhängige Variablen (microRNAs, SNPs, Proteine, Metaboliten) nutzen, um eine Funktion zu erstellen, die schließlich die abhängige Variable (bestimmter Phänotyp) modelliert.

„Durch Deep Learning erkennt der Computer, ob aus einer Blutvorläuferzelle ein rotes oder ein weißes Blutkörperchen entsteht.“

Wir haben mithilfe von Regressionsmodellen zum Beispiel einen genetischen Risikoscore für Typ-1 Diabetes erstellt. Hierzu bildeten wir zunächst aus einem Datensatz von über 4.500 Proben einen Risikoscore aus der Kombination von SNPs in zwölf Diabetes-Typ-1 assoziierten Genen, indem wir pro Patient die Anzahl der Risikoallele aufsummierten. Anschließend gewichteten wir den Beitrag jedes SNPs mit Hilfe von Regressionsmodellen. Die Vorhersage wurde hierdurch erheblich präziser, da dieses Modell variierende Effektstärken verschiedener SNPs abbildet. Diese Arbeiten dienen unter anderem als Basis für ein bayernweites Screening zur Früherkennung von Diabetes Typ 1 bei Kleinkindern (Fr1Da-Studie).

In der Kategorie unüberwachtes Lernen finden sich zum Beispiel sogenannteClustering-Methoden. Bei diesen sucht man mit Cluster-Verfahren wie k-Means in großen Datensätzen nach Clustern und ordnet die Objekte diesen Clustern zu.

Wir verwenden Clustering-Methoden unter anderem bei der Analyse hochdimensionaler Large-scale Sequenzierungsdaten (RNA oder qPCR-Daten), die sich nur schwer verarbeiten lassen. Hierzu setzen wir sogenannte Diffusion Maps ein, die Strukturen in den Daten aufspüren und zugrundeliegende Prozesse visualisieren. Bei dieser Technik werden multidimensionale Daten mit bis zu 10.000 Dimensionen auf zwei bis drei Dimensionen projiziert. Die Struktur der Daten bleibt hierbei möglichst gut erhalten. Nichtlineare Dimensionsreduktionsverfahren, wie Diffusion Maps, sind in Big Data-Anwendungen unerlässlich, um die Daten auf das eigentliche Modell-Lernen vorzubereiten oder zu visualisieren.

Eine weitere Anwendung maschinellen Lernens sind künstliche neuronale Netze. Diese bestehen aus geschichteten Operationseinheiten (Neuronen), die Signale aus verschiedenen Richtungen aufnehmen, verarbeiten und aufaddieren, um daraus einen Output zu generieren. Neuronale Netze haben eine lange Geschichte, die bis in die 40er Jahre zurückreicht. Sie verschwanden jedoch zu Beginn des 21. Jahrhunderts von der Bildfläche, weil sowohl die Leistung der Computer als auch die Datenmenge für ihren Einsatz nicht ausreichten.

Durch die heutige Rechenleistung und den Zuwachs an Big Data hat sich diese Situation jedoch geändert. Neuronale Netze sind inzwischen für viele Forschungsfelder interessant. Wir benutzen sie beispielsweise für die Computervision, bei der unter anderem Mikroskopie-Daten anhand von Algorithmen klassifiziert werden. Hier sind insbesondere das „Deep Learning“ oder „Deep Neural Networks“ hervorzuheben. Hinter diesen Begriffen verbergen sich neuronale Netze mit sehr vielen Schichten, die sich aber dank weiterentwickelten Algorithmen immer noch effizient trainieren lassen.

Mit dieser Methode kann man einen Computer zum Beispiel so weit trainieren, dass er das Stadium oder den Typ einer Zelle anhand ihrer Morphologie erkennt. Ein Anwendungsbeispiel sind Time-Lapse Mikroskopie-Daten während der Hämatopoese. Durch Deep Learning erkennt der Computer, ob aus einer Blutvorläuferzelle ein rotes oder ein weißes Blutkörperchen entsteht. Im Labor ist dies nur mit zeitlich aufwendigeren Methoden möglich, zum Beispiel über Oberflächen-Marker. Auch Image-basierte Durchflusszytometrie-Daten lassen sich mit Machine-Learning-Algorithmen auswerten. In einem konkreten Beispiel brachten wir dem Computer bei, den Status des Zellzyklus anhand der Zellform zu ermitteln. Mit diesem Label-freien Assay umgeht man aufwendige Fluoreszenzfärbungen.

Aus diesen Beispielen sind die Vorteile systembiologischer Methoden zur Lösung biologischer Fragestellungen ersichtlich: Zum einen sparen sie enorme Kosten, da

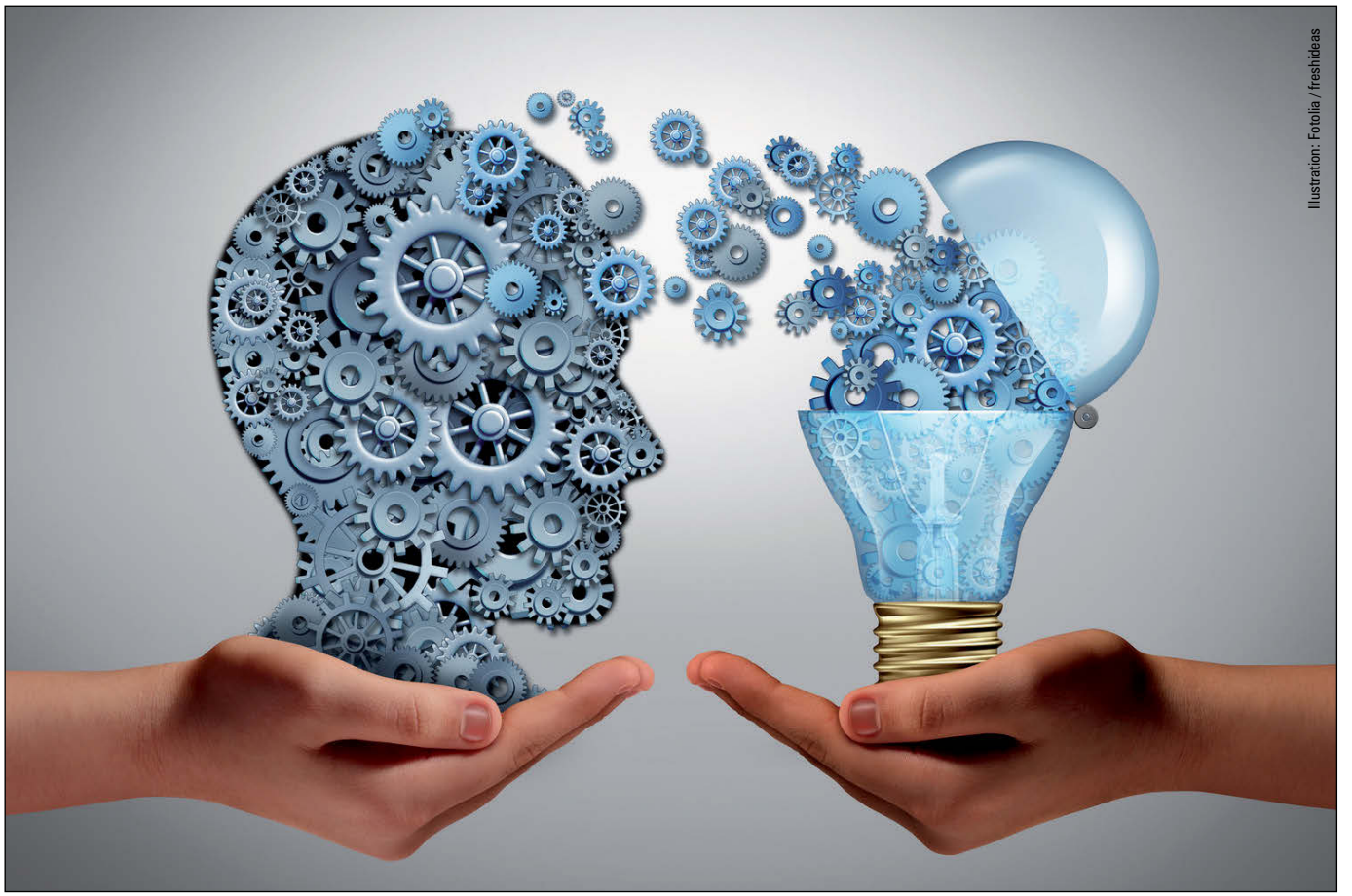

Illustration: Fotolia / freshideas

gerade zellbiologische Techniken, wie Labelling oder Antikörper-basierte Assays sehr teuer und zeitaufwendig sind. Zum anderen können die hierdurch freigewordenen Fluoreszenzkanäle für andere Fragestellungen genutzt werden.

Besonders vielversprechend ist das maschinelle Lernen bei der Einzelzellanalyse. Bei dieser werden aus einer heterogenen Zellpopulation individuelle Zellprofile bestimmt – jede Zelle wird also einzeln betrachtet. Die Analyse einzelner Zelltypen ist in vielen Anwendungen essentiell: Krebszellen, Differenzierungsprozesse oder Krankheitsentstehung lassen sich nur anhand klar definierter, detaillierter Zellprofile verstehen. Klassische Strategien, in denen Subpopulationen wie beispielsweise krebsartige Zellen in der großen Zahl normaler Zellen verschwinden, sind hier wenig hilfreich.

Diese Analysen sind jedoch aufgrund der notwendigen Amplifikation, dem Umgang mit kleinsten Assay-Mengen oder der schwierigen Isolation einzelner Zellen, mit großen Störfaktoren verbunden. Zudem erschweren Veränderungen der Genexpression, unterschiedliche Zellzyklus-Stadien oder variierende Differenzierungsgrade einzelner Zellen den Vergleich. Durch die Kombination von Einzelzellanalysen und statistischen Modellen lassen sich solche Unsicherheitsfaktoren aber herausrechnen und man erhält ein genaueres Abbild des Zelltyps.

Mithilfe des sogenannten Single-Cell Latent Variable-Modell (scLVM) gelang es uns beispielsweise, unterschiedliche, ansonsten nicht nachweisbare Reifestadien von T-Zellen in ihrer Entwicklung hin zu Th2-Zellen zu detektieren und zu charakterisieren. Bei dieser Methode schätzen wir zunächst mithilfe von Zellzyklus-Transkripten den Status der einzelnen Zellen im Zellzyklus ab. Anschließend wird dieser durch ein Regressionsmodell korrigiert, so dass die Zellen vergleichbar sind.

Die hier beschriebenen Beispiele zeigen nur einen Bruchteil der Möglichkeiten, die Mathematik und Informatik bei der Lösung biologischer Fragen bieten. Spannende neue Themen sind vor allem im Umfeld von Big Data zu finden. Etwa bei der Aufteilung von Patientengruppen (Stratifizierung) aufgrund von Multi-omics Daten (Präzisionsmedizin) oder bei der systematischen Analyse bildgebender Daten.

Auch methodisch sind neue Ansätze gefragt. Mit Transfer-Learning-Techniken lassen sich zum Beispiel heterogene Daten aus verschiedenen Domänen integrieren, während das systematische Hochskalieren der Algorithmen den Umgang mit großen Datensätzen erleichtert.

Obwohl das Thema maschinelles Lernen zunächst eher trocken klingt, ist es gerade in der heutigen Zeit, in der die technische Entwicklung sprunghaft vorangeht, das genaue Gegenteil. Nicht umsonst bezeichnet der *Harvard Business Review* die Arbeit des Data Scientist als „Sexiest Job in the 21st Century".

Anna Sacher *ist Science Manager am Institute of Computational Biology (ICB) des Helmholtz Zentrum München.*
Fabian Theis *ist Direktor des ICB und leitet die Gruppe Machine Learning.*