Hananeh Aliee, Anna Sacher, Fabian J. Theis

4. Analyse von Einzelzellgenomik-Daten mit Methoden des maschinellen Lernens

In:
Walter, Jörn / Schickl, Hannah (Hrsg.): Einzelzellanalyse in Forschung und Medizin : eine Stellungnahme der interdisziplinären Arbeitsgruppe Gentechnologiebericht.
ISBN: 978-3-939818-84-7
Berlin-Brandenburgische Akademie der Wissenschaften, 2019. S. 48-54

Persistent Identifier: urn:nbn:de:kobv:b4-opus4-32821

Hananeh Aliee, Anna Sacher, Fabian J. Theis

4. ANALYSE VON EINZELZELLGENOMIK-DATEN MIT METHODEN DES MASCHINELLEN LERNENS

4.1 EINFÜHRUNG

Sowohl in der Wissenschaft als auch in der Industrie werden Datensätze stetig größer, da immer mehr moderne Geräte und Technologien sogenannte *Big Data* generieren und sammeln. „Big Data" ist ein Begriff für große und komplexe, oft unstrukturierte Datenmengen, die eine Bearbeitung mit rechnerbasierten Hilfsmitteln erforderlich machen, um bedeutsame Informationen daraus ziehen zu können. Die Analyse solcher Daten wird als „Data Science" (Datenwissenschaft) bezeichnet. Sie eröffnet neue Möglichkeiten der Kombination von Daten aus unterschiedlichen Quellen und ermöglicht so tiefere Einblicke in ein spezifisches Problem für eine bessere Entscheidungsfindung.

Zur Extraktion von Werten aus Daten kommen oft Methoden des *maschinellen Lernens* (ML) zum Einsatz, eine der großen Triebfedern der Big-Data-Revolution. ML ist ein Unterbegriff der *künstlichen Intelligenz* (KI), mit der im Allgemeinen menschliche Intelligenz für bestimmte Aufgaben nachgeahmt werden soll. Genauer bezeichnet ML unterschiedliche Berechnungsalgorithmen für autonomes Lernen aus gekennzeichneten und ungekennzeichneten Daten[1], um datengestützte Einblicke zu liefern und Entscheidungsfindungen sowie Vorhersagen zu unterstützen. Durch die immer weiter anwachsende Datenflut kann es allerdings sein, dass konventionelles ML nicht mehr leistungsstark genug ist, um die Komplexität innerhalb der Daten richtig zu erfassen. Aus diesem Grund entstand *Deep Learning* (DL) als neuer Bereich des ML. Deep Learning ist eine ML-Technik basierend auf künstlichen neuronalen Netzwerken, die einfache nicht-lineare Verarbeitungseinheiten[2] in mehreren Schichten verknüpfen. Die Architektur von Deep-Learning-Modellen kann komplizierte, hierarchische, statistische Muster innerhalb von Datensätzen überwacht („supervised", beispielsweise zur

1 Ungekennzeichnete Daten („unlabeled data") sind Daten ohne zugehörige Informationen, wohingegen gekennzeichnete Daten („labeled data") zusätzliche Informationen über diese Daten umfassen.

2 Diese Einheiten, die man als *künstliche Neuronen* bezeichnet, bilden grob die Neuronen in einem biologischen Gehirn nach. Über eine Verbindung kann ein Signal von einem künstlichen Neuron an ein anderes weitergegeben werden. Das empfangende Neuron verarbeitet das Eingangssignal und leitet es an andere verbundene künstliche Neuronen weiter.

Klassifizierung) oder nicht überwacht („unsupervised", z. B. für Gruppierungen) erfassen.[3] Der Hauptvorteil von DL-Algorithmen ist, dass sie schrittweise immer komplexere Merkmale aus Daten lernen können. Diese Merkmalsauswahl benötigt kein Expertenwissen für das vorliegende Problem, allerdings sind für gewöhnlich auch größere Datensätze erforderlich.

DL hat in den letzten Jahren viele Bereiche wie *Computervision* (zum maschinellen Verständnis digitaler Bilder und Videos) und *Natural Language Processing* (zum maschinellen Verständnis natürlicher Sprache) revolutioniert und breite Anwendungsgebiete u. a. in der Astronomie, Robotik, im Finanz- und Gesundheitswesen gefunden. In diesem Kapitel liegt der Fokus auf der Gesundheitsforschung, insbesondere auf der Genomik. Dieser Bereich hat in den letzten Jahren durch neue Fortschritte in biomedizinischen Techniken wie *Next-Generation-Sequencing* (NGS), die heute routinemäßig riesige Mengen genomischer Daten generieren, ein exponentielles Wachstum gezeigt. Bei NGS-basierten Technologien wie Genomik, Transkriptomik, Proteomik und Epigenomik[4] liegt der Fokus immer stärker auf der Profilerstellung einzelner Zellen. Im Gegensatz zu traditionellen Profilierungsmethoden zur Beurteilung großer Zellpopulationen werden bei Einzelzelltechnologien einzelne Zellen isoliert, um zellspezifische Sequenzierungsbibliotheken, eine Sammlung ähnlicher Moleküle, zu erstellen. Dabei wird jede Zelle individuell mit einem zellspezifischen molekularen Barcode markiert. Einzelzelltechnologien ermöglichen dann eine Profilierung der Informationen über Tausende bis Millionen einzelner Zellen in einem einzigen Experiment. Dadurch ist es inzwischen möglich geworden, die Heterogenität auch zwischen ähnlichen Zelltypen aufzudecken (siehe Aschenbrenner, Mass, Schultze, Kapitel 3) und potenziell komplexe und seltene Zellpopulationen, Zelldynamiken, regulatorische Beziehungen zwischen Genen und sog. Entwicklungstrajektorien[5] bestimmter Zelllinien zu erkennen (Hwang et al., 2018; siehe Junker, Popp, Rajewsky, Kapitel 2). Die Komplexität der Einzelzelldaten gepaart mit ihrem

3 „Supervised" bedeutet, dass Funktionen auf der Basis bekannter Datensätze abgeleitet werden, die die Klassifizierung unbekannter Daten ermöglichen. „Unsupervised" bedeutet, dass unbekannte Daten untersucht und Strukturen innerhalb dieser Daten identifiziert werden, was dann Gruppierungen ermöglicht.

4 In der Genomik geht es um die Untersuchung des gesamten Genoms, in der Transkriptomik um die Untersuchung aller Transkripte von Genen (Genexpressionsprodukte, RNA), in der Proteomik wird die Gesamtheit aller Proteine untersucht, und in der Epigenomik alle epigenetischen Daten innerhalb von Zellen. Siehe Walter/Gasparoni, Kapitel 1.

5 Eine Entwicklungstrajektorie bezeichnet den Verlauf einer Entwicklung über die Zeit hinweg. Eine solche Entwicklungstrajektorie kann aus Einzelzelldaten rekonstruiert werden, wenn man Zellen zu vielen einzelnen Zeitpunkten vermisst und sie dann entlang einer virtuellen Zeitachse anordnet (pseudotemporale Trajektorie).

gigantischen Umfang macht die Einzelzellanalyse zu einem paradigmatischen Fall von Big Data. Daraus ergibt sich die Notwendigkeit der Entwicklung von Analysemethoden, die große Datensätze über eine große Anzahl von Zellen verarbeiten können. Dieses Kapitel liefert einen Überblick über die Einzelzell-Transkriptomik als eine der populärsten Einzelzelltechnologien und deren Herausforderungen und Möglichkeiten aus der Perspektive der modernen Analytik basierend auf ML und DL.

4.2 MASCHINELLES LERNEN IN DER EINZELZELL-TRANSKRIPTOMIK

Die Einzelzell-RNA-Sequenzierung (Single-cell-RNA-Sequencing, scRNA-seq) umfasst die Profilierung aller Messenger-RNA (mRNA)[6] in einer einzelnen Zelle und liefert das Genexpressionsprofil von Hunderttausenden und sogar Millionen individueller Zellen. Damit erhebt scRNA-seq Big Data mit überlegener statistischer Aussagekraft, die neue Möglichkeiten der Anwendung von maschinellem Lernen und Deep Learning für die Einzelzelldatenanalyse eröffnen.

Da sich die Messungen einzelner Zellen im Femtoliterbereich bewegen, sind technische Störfaktoren erhöht und die erhaltenen Expressionswerte oft verrauscht.[7] Daraus ergeben sich mehrere Herausforderungen für die Berechnung und Statistik in der Erkennung von Mustern in der Genexpression, zum Beispiel von Zelltypen. Häufig wird eine zusätzliche Qualitätskontrolle durchgeführt, um fehlerhafte Messungen auszuschließen, verursacht zum Beispiel durch sogenannte Ausreißer („drop-out effect") oder mögliche Dubletten,[8] gefolgt von einer *Normalisierung,*[9] bei der zellspezifische Unterschiede in der Sequenzierung und andere technische Störfaktoren korrigiert werden. Anschließend wird eine *Merkmalsauswahl*[10] und

6 Messenger-RNA werden vom Genom der Zelle abgeschrieben und am Ribosom in Proteine übersetzt. Sie stellen damit das aktive Genom dar.

7 „Rauschen" bedeutet in diesem Zusammenhang, dass die Daten in Bezug auf die zu untersuchende Frage irrelevante oder zufällige Signale enthalten, die zunächst herausgefiltert werden müssen, um signifikante Signale identifizieren zu können.

8 Als „Ausreißer" bezeichnet man Zellen, die vom durchschnittlichen Expressionsgrad ihres Zelltyps abweichen und somit die Identifizierung gewöhnlich exprimierter Gene erschweren. Dubletten sind Expressionsprofile, die versehentlich aus zwei und nicht aus einer Zelle generiert werden, oft infolge von Fehlern beim Sortieren oder Erfassen der Zellen. Sie können die korrekte Interpretation der Ergebnisse beeinträchtigen, z. B. indem sie auf das Vorhandensein von Zwischenpopulationen oder Übergangsstadien hinweisen, die es in Wahrheit gar nicht gibt.

9 Bei der Normalisierung werden Auszählungsdaten skaliert, um eine korrekte relative Fülle von Genexpressionen zwischen verschiedenen Zellen zu erhalten.

10 Bei der Merkmalsauswahl wird der Datensatz gefiltert, um nur Merkmale/Variablen (in diesem Zusammenhang: Gene) zu behalten, die etwas über die Variabilität innerhalb der Daten aussagen.

eine *Dimensionsreduktion*[11] durchgeführt, wobei die am meisten informativen Gene und die stärksten Signale aus dem Hintergrundrauschen herausgefiltert werden (Luecken/Theis, 2019). Die Qualitätskontrolle ist für die nachgeschaltete Analyse, zum Beispiel die Gruppierung von Zellen zur Erkennung von Teilpopulationen oder zur Ableitung künftiger Zellentwicklungen, erforderlich. Im Folgenden werden einige Anwendungsgebiete überwachter und nicht überwachter Lerntechniken in der Downstream-Analyse transkriptomischer Daten erörtert.

Überwachtes Lernen in der Einzelzell-Transkriptomik

Das überwachte Lernen („supervised learning") ist ein Bereich des maschinellen Lernens, der Training mit bekannten Daten erfordert, sodass aus diesen gekennzeichneten Daten eine Funktion abgeleitet werden kann, die sich zur Kartierung von ungekennzeichneten Daten und Ausgangsvariablen eignet. Ein überwachtes Lernmodell wird zunächst mit einem *Trainingssatz* aus Eingangszielpaaren trainiert, damit es die Modellparameter lernt. Um messen zu können, wie gut eine Funktion zum Trainingssatz passt, wird eine Verlustfunktion definiert, um Fehler in der Vorhersage zu bestimmen. Dadurch sollen die Modellparameter durch Minimierung der Vorhersagefehler optimiert werden. Außerdem wird das Modell mit einem bestimmten *Validierungssatz* verifiziert und anschließend die Leistung der abgeleiteten Funktion mit einem vom Trainingssatz separaten *Testsatz* beurteilt. Die Genauigkeit der Vorhersagen wird mit unterschiedlichen Beurteilungsmessgrößen wie zum Beispiel dem Pearson-Korrelationskoeffizienten gemessen. Die Hauptanwendungsbereiche von überwachtem Lernen sind Klassifizierung und Regression.[12]

Im Bereich der Einzelzell-Transkriptomik wird überwachtes Lernen vor allem zur *Zellannotation* verwendet. Hier werden unbekannten Zellen Zelltypen anhand einer Reihe von Referenzdatensätzen mit gekennzeichneten Zelltypen zugeordnet. Dies entspricht im Bereich der Genomikdaten in etwa dem routinemäßigen Einsatz der Durchflusszytometrie[13] zur Diagnose von Erkrankungen wie Blutkrebs. Konventionell wurden Zellen basierend auf einer Reihe von Markern annotiert,

11 Dimensionsreduktion ist ein Verfahren zur Senkung der Anzahl von Zufallsvariablen durch Identifizierung einer gewissen Anzahl relevanter Variablen.

12 Regression ist eine Reihe statistischer Verfahren zur Einschätzung des Verhältnisses zwischen den Variablen.

13 Durchflusszytometrie ermöglicht das Bestimmen von molekularen und physikalischen Eigenschaften von Zellen mithilfe von Fluoreszenz-markierten Proben wie z. B. Antikörpern.

was arbeitsaufwendig ist und eine umfassende Literaturübersicht über zelltypspezifische Gene erforderlich macht. Außerdem variieren diese Gene oft zwischen einzelnen experimentellen Gegebenheiten, was den Vergleich der Ergebnisse erschwert (Pliner et al., 2019). Klassische Techniken des überwachten Lernens sind besser, da sie aus den gekennzeichneten Daten automatisch wichtige Merkmale (oder Gene) erfassen, was eine exaktere Zellannotation und weniger Diskrepanzen zwischen den Klassifizierungen verschiedener Experimente ermöglicht. In der rechnerbasierten Analyse werden zahlreiche Klassifizierungsmodelle wie zum Beispiel die logistische Regression, *Support Vector Machines* und *Random Forests* eingesetzt. Allerdings werden bei immer größerem Datenvolumen Deep-Learning-Modelle bei Zellannotationsaufgaben gegebenenfalls den klassischen Machine-Learning-Modellen vorgezogen.

Unüberwachtes Lernen in der Einzelzell-Transkriptomik

Beim unüberwachten Lernen („unsupervised learning") werden nützliche Strukturen oder Muster aus nicht gekennzeichneten Datensätzen abgeleitet. Klassisch kommen nicht überwachte Lernalgorithmen zur Gruppierung von Daten, für die Dimensionsreduktion und zur Visualisierung und Einbettung zum Einsatz. Künstliche neuronale Netzwerke können einige dieser Ansätze generalisieren. *Autoencoder* komprimieren die Daten beispielsweise in einen niedrigdimensionalen Code und dekomprimieren diesen dann zur Rekonstruktion der ursprünglichen Eingangsdaten. Ein Autoencoder ermöglicht aber nur die ungefähre Kopie der Eingangsdaten in den Output, was das Modell dazu zwingt, eine Dimensionsreduktion durchzuführen, indem es lernt, Rauschen zu ignorieren. Im Bereich der Einzelzell-Transkriptomik kommen Autoencoder für Entrauschen sowie zur Dimensionsreduktion zum Einsatz. Einbettungstechniken wie t-SNE können zur Kartierung der komprimierten Daten in einer 2D-Ebene angewendet werden. Spezifische Geräuschmerkmale der scRNA-seq-Daten wie zum Beispiel *Zero-Inflated Negative Binomial* (ZINB) können ebenfalls mit maßgeschneiderten Verlustfunktionen innerhalb des Autoencoder-Rahmens behandelt werden (Eraslan et al., 2019).

Eine weitere leistungsstarke Anwendung von nicht überwachtem Lernen ist die *Clusteranalyse* zur Definition von Zelltypen innerhalb der scRNA-seq-Daten. Das Ziel der Clusteranalyse ist die Gruppierung von Zellen basierend auf Ähnlichkeiten in deren Genexpressionsprofilen. Die Clusteranalyse ist die

Basis mehrerer Atlasprojekte, insbesondere des Human Cell Atlas.[14] Bei solchen Forschungsprojekten werden mehrere Einzelzelldatensätze in einen Atlas aufgenommen, der umfassende Referenzkarten aller menschlichen Zellen enthält. Damit ein Zellatlas einen bestimmten Zweck erfüllen kann, ist eine der wichtigsten rechnerischen Herausforderungen die Entwicklung zuverlässiger Methoden zur unüberwachten Gruppierung der Zellen (Kiselev et al., 2019).

Die Zelldiversität wird allerdings in einem einzelnen Klassifizierungssystem wie dem Clustering nicht ausreichend beschrieben. Immerhin sind die biologischen Prozesse, die zur Entwicklung der beobachteten Heterogenität führen, kontinuierliche Prozesse. Um also Übergänge zwischen Zelltypen, Prozesse der verzweigten Differenzierung oder graduelle, nicht synchronisierte Veränderungen in der biologischen Funktion erfassen zu können, benötigen wir dynamische Modelle der Genexpression. Diese Methodenklasse wird als *Trajectory Inference* bezeichnet. In der Trajektorienanalyse (siehe Junker, Popp, Rajewsky, Kapitel 2) werden die Daten als Schnappschuss eines dynamischen Prozesses bzw. auf einer virtuellen Zeitachse abgebildet. Die Zellen werden dann entlang einer solchen virtuellen Zeitachse geordnet und durch eine kontinuierliche Variable beschrieben, die man als *Pseudozeit* bezeichnet. Die Pseudozeitanalyse basiert oft auf der transkriptionellen Distanz von Zellen zu einer Ursprungszelle und beschreibt die Entwicklung als Übergang im transkriptomischen Zustand (d. h. Trajektorie), und nicht als Übergang in Echtzeit. Die pseudotemporale Ordnung von Zellen hilft dabei, zu verstehen, wie sich Zelltyphäufigkeiten als Reaktion auf Entwicklungs- oder Umweltsignale verändern, die physiologischen Mechanismen von Gesundheit und Krankheit unterliegen. Beispielsweise wird damit festgestellt, wie die Häufigkeit eines bestimmten Zelltyps im Laufe eines Prozesses abnimmt, indem die Sterblichkeitsrate steigt oder eine Differenzierung zu anderen Zelltypen stattfindet. Es ist dabei wichtig, die Ursache einer solchen Verschiebung zu verstehen, insbesondere wenn der Prozess im Zusammenhang mit einer Erkrankung steht.[15] Eine weitere interessante Frage, die eine Pseudozeitanalyse beantworten kann, ist die Frage, wie sich Stammzellen oder Vorläuferzellen differenzieren, um ein Organ aus unterschiedlichen Zelltypen auszubilden. Um die allgemeine Topologie der Daten zu verstehen und Trajektorien abzuleiten, kommen häufig nicht-lineare Methoden der Dimensionsreduktion (z. B. *Manifold Learning*) zum Einsatz (Wolf et al., 2019).

14 Siehe: https://www.humancellatlas.org/ [21.06.2019] und Walter/Gasparoni, Kapitel 1.

15 Ein Beispiel hierfür wäre der Zusammenhang zwischen einem Rückgang in der Häufigkeit von Betazellen der Bauchspeicheldrüse und Diabetes.

4.3 AUSBLICK

Die Einzelzell-RNA-Sequenzierung ist eine leistungsstarke Methode zur Entdeckung interzellulärer Heterogenität. Dabei liegt der Fokus auf der Charakterisierung individueller Zellen. Dadurch können komplexe und seltene Zellpopulationen identifiziert werden, ebenso wie regulatorische Beziehungen zwischen Genen und die Trajektorien bestimmter Zelllinien in der Entwicklung. Mehrere Studien haben gezeigt, wie nützlich scRNA-seq Technologien insbesondere zur Untersuchung der frühen embryonalen Entwicklung sowie zur Erkennung der Komplexität von Krebs und anderen Erkrankungen sind. Sowohl die Komplexität als auch der gigantische Umfang der Einzelzelldaten stellen große Herausforderungen in der Datenanalyse dar. Dazu kommt, dass es sich bei der Einzelzellanalyse um ein neues Forschungsfeld handelt, für das standardisierte Analysemethoden erst noch entwickelt werden müssen.

4.4 LITERATUR

Eraslan, G. et al. (2019): Single-cell RNA-seq denoising using a deep count autoencoder. In: Nat. Commun 10(1): 390.

Hwang, B. et al. (2018): Single-cell RNA sequencing technologies and bioinformatics pipelines. In: Exp. Mol. Med. 50(8): 96.

Kiselev, V. Y. et al. (2019): Challenges in unsupervised clustering of single-cell RNA-seq data. In: Nat. Rev. Genet 20(5): 273–282.

Luecken, M. D./Theis, F. J. (2019): Current best practices in single-cell RNA-seq analysis: a tutorial. In: Mol Syst Biol 15: e8746.

Pliner, H. A. et al. (2019): Supervised classification enables rapid annotation of cell atlases. In: bioRxiv, Online-Publikation 04.02.2019. DOI: 10.1101/538652.

Wolf, F. A. et al. (2019): PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. In: Genome Biol 20(1): 59.