Big Data

Mit Klassifikatoren und multimodaler Datenfusion Big Data Sinnvolles entlocken

Mit großen Datenmengen kennt sich Prof. Hans A. Kestler aus. Er leitet das Institut für Medizinische Systembiologie an der Universität Ulm und kann sich vor Kooperationsanfragen der Kliniker kaum retten. Walter Pytlik hat ihn für die BIOPRO gefragt, ob die Voraussetzungen für eine verstärkte Nutzung von Big Data in der biomedizinischen Forschung schon ausreichen.

Prof. Hans A. Kestler
© Universität Ulm

Wo sehen Sie die größten Herausforderungen auf dem Weg zu einem vernetzten Gesundheitssystem.

Das ist ein kontinuierlicher Prozess, gewissermaßen geht es um fortlaufende Verbesserung. Die größten Herausforderungen sehe ich im Datenschutz, in vielen rechtlichen Aspekten, zum Beispiel wem die Daten gehören und auf der technischen Seite. Auf der semantischen Seite liegt die größte Herausforderung, Daten unterschiedlicher Standorte, möglicherweise auch innerhalb der gleichen Klinik, vergleichbar zu machen.

Beginnen wir beim Datenschutz.

Da bin ich kein Experte. Aber ich sehe beispielsweise Probleme bei der länderübergreifenden Vernetzung, da hier viele Akteure mitsprechen, Kostenträger, Kliniken oder Datenschutzbeauftragte. Vorstellbar sind Bedenken, die Daten zu integrieren, nicht nur aus Datenschutzgründen.

Was meinen Sie damit?

Wollen wir diese Datenintegration überhaupt, wenn sich damit Rückschlüsse auf die einzelne Klinik oder die Abteilungen ergeben? Ein Beispiel: aus der Integration von Behandlungsdaten werden möglicherweise unterschiedliche Behandlungsarten erkennbar. Es könnte sich ergeben, dass manchmal nicht leitliniengemäß therapiert wird. Das kann indirekt zu einer vielleicht trügerischen Vergleichbarkeit von Behandlungen führen, die im Detail vielleicht gar nicht gegeben ist, da die Patientenkollektive unterschiedlicher Kliniken unterschiedlich strukturiert sein können.

Wenn die Zusammenführung unterschiedlicher Daten dazu führt, dass Rückschlüsse auf die Leistungsfähigkeit, die Behandlungskollektive oder alle möglichen leistungsrelevanten Parameter gefällt werden, dann könnte es schwierig werden, sie umzusetzen. Hier ist dann wohl der Gesetzgeber stark gefordert. Gut umgesetzt werden wird das erst, wenn Gesetze erlassen werden, die die Kliniken oder die Kostenträger verpflichten, gewisse Daten unter bestimmten Voraussetzungen zur Verfügung zu stellen. Das Problem liegt im Detail. Wenn das im Detail nicht gut gelöst ist und alle Beteiligten nicht an einem Strang ziehen, dann kann man jede Zusammenführung torpedieren.

Nichtdestotrotz wird diese Zusammenführung früher oder später kommen, muss kommen. Die Chancen für eine Behandlungsverbesserung sind einfach zu groß, beispielsweise durch die Identifikation neuer molekularer Subgruppen, gerade bei seltenen Erkrankungen kann dies - und wird ja bereits durchgeführt - zu großen Behandlungserfolgen führen, weil man an einem einzelnen Standort nicht genügend Patientendaten hat.

Die Daten gehören dem Patienten?

Ja – aber nicht alle Forscher sind damit derzeit so befasst, weil oft mit anonymisierten oder pseudonymisierten Daten gearbeitet wird. Das wird sich ändern, wenn zukünftig mehr intelligente Verfahren in der Klinik Einzug halten wie Künstliche Intelligenz und Entscheidungsunterstützung. De facto gibt es die bereits jetzt schon, wie Deep Learning, künstliche neuronale Netze oder andere maschinelle Lernverfahren.

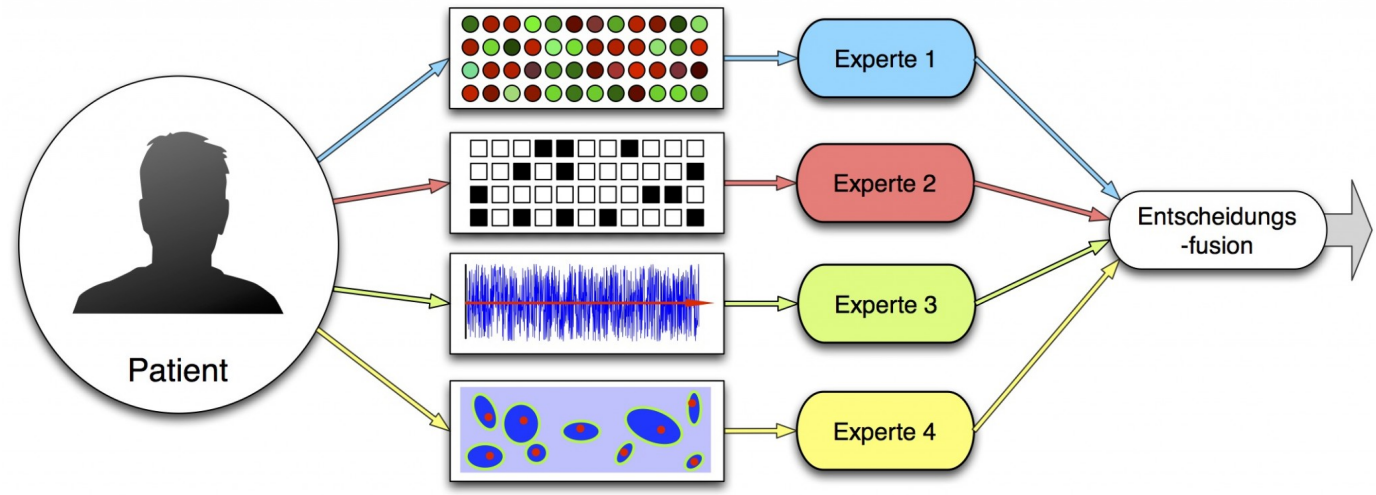
Was muss mit den Daten passieren, damit bei Einsatz genannter Methoden Sinnvolles entsteht, damit aus Big Data Smart Data werden?

Meine Arbeitsgruppe und ich sind stark an interpretierbaren Entscheidungsverfahren interessiert und wollen diese automatisch generieren, also interpretierbare Klassifikation. Wir wollen keine Blackbox, wie es oft bei Deep Learning der Fall ist. Wir wollen einfache, gut generalisierende Klassifikatoren bauen, aus deren Struktur sich lernen lässt, so dass man zum Beispiel weiß, welcher molekulare Signalpfad wichtig für eine Einstufung ist oder welche klinischen Daten wichtig für eine Therapieentscheidung sind. Kurz, wir bauen Klassifikatoren, die möglichst gut auf neue Daten verallgemeinern können.

Wie geht das vonstatten?

Ein solcher Klassifikationsprozess läuft so ab: Wir trainieren mit vorhandenen Daten, adaptieren ein System und wenden dieses System auf neue Daten an. Wir wollen beispielsweise von der Klassifikation einer Tumorentität auf die einer anderen schließen. Das ist wiederum ein Subaspekt der Forschung, also Klassifikatoren, die voneinander lernen.

In unseren Klassifikationsansätzen bauen wir Tumorboards gewissermaßen nach, also interdisziplinäre Tumorkonferenzen, wie sie in onkologischen Spitzenzentren üblich sind. Wir haben eine Mixtur algorithmischer Experten. Der eine Experte schaut sich die Daten nur unter



Schematische Darstellung der multimodalen Datenfusion. Damit lässt sich Wissen aus unterschiedlichen Abstraktionsebenen miteinander verbinden. Ziel ist es, die Komplexität des Klassifikators zu verkleinern und damit dessen Generalisierungsleistung zu erhöhen.

einem gewissen Gesichtspunkt an, der nächste nur unter einem anderen. Danach bilden sie eine Konsensus-Entscheidung. Solche Vorgehensweisen bilden wir algorithmisch ab. Darin besteht großenteils unsere Forschung. Weiß man, dass nur eine Untermenge von Experten für die Entscheidungsfindung wichtig ist, erlaubt das Rückschlüsse auf die Genese der Erkrankung.

Ein weiterer Ansatz ist die multimodale Datenfusion. Dabei werden Daten aus unterschiedlichen Modalitäten (siehe Abbildung) zusammengeführt. Damit lässt sich Wissen aus unterschiedlichen Abstraktionsebenen miteinander verbinden. Die Herausforderung ist, dass das Wissen in unterschiedlichen Ebenen repräsentiert ist. Beispiel: Ein Gen ist Teil eines oder mehrerer Signalwege. Diese Information kann uns helfen, bei einer RNA-Sequenzierung die Komplexität des Klassifikators zu verkleinern und damit dessen Generalisierungsleistung zu erhöhen. Das ist das, was wir mit dieser multimodalen Datenfusion machen.

Davon erhoffen wir uns eine Reduktion von Komplexität und dadurch eine verbesserte Verallgemeinerungsfähigkeit. Letztlich läuft es immer darauf hinaus, Verfahren zu entwickeln, die möglichst gut verallgemeinern können, das heißt möglichst gut auf neue Daten antworten können, auch wenn wir diese nie zum Training verwendet haben.

Gibt es einen Goldstandard der medizinischen Dokumentation, von den Laborwerten bis zu Sequenzdaten?

Nein, den einen Standard gibt es nicht, es gibt viele. Das ist einerseits schlecht, andererseits aber nicht. So begrüßenswert jeweils ein Standard wäre, er würde auch einschränken. Denn auf dem Gebiet der Technologie ist vieles stark im Fluss. Würde man vieles festzurren, würde es damit auch nicht unbedingt eine gute Weiterentwicklung geben. Trotz einer neueren, besseren Technologie, wäre man auf die alte im Standard festgelegt.

Es gibt natürlich verschiedene Standardisierungsversuche, auch in der Labordiagnostik, beispielsweise. Allerdings werden diese nicht von allen angewandt. Und es gibt immer wieder Gründe, diese nicht zu verwenden. Die Herausforderung wird bestehen bleiben. Minimale Werte wie die Daten auf der Versichertenkarte lassen sich leicht zusammenbringen, aber bei gewissen Blutwerten wird allein schon vielleicht die Labordiagnostik unterschiedlich sein.

Das setzt aber voraus, dass die Daten mit möglichst viel konzisen und exakten Metadaten versehen werden.

Viele Metadaten liegen bereits vor, es kommt hier sehr auf die Fragestellung an. Sequenzdaten mit Zusatzinformationen zu versehen ist beispielsweise über allgemein verfügbare Datenbanken wie KEGG* möglich. Das ist allgemeines Wissen, was sich laufend ändert und hilft, Daten zu strukturieren, zu gruppieren. Selbst dieses abstrakte Wissen können wir versuchen, in einen solchen Klassifikationsprozess einzubeziehen. In unserer Forschung zumindest hat das sehr gute Ergebnisse gebracht.

Sie geben den Daten also Struktur, Ontologien wie es im Fachjargon wohl heißt...

Ontologien sind eine weitere Möglichkeit. Das kann noch viel weiter gehen, indem man genaue Signalnetzwerke kennt. Da ist noch viel mehr Struktur in den Daten. Wie viel man von dieser Struktur nehmen kann, um sie in diese Prozesse miteinzubeziehen, das ist auch Gegenstand aktueller Forschung.

Es gibt also bereits Tools, Methoden, mit denen man den Daten Sinnvolles entlocken, Korrelationen herstellen kann?

Das Beispiel der Ontologien und Pathways sind keine Korrelationen. Da sind Gene gewissen Begrifflichkeiten zugeordnet und diese letzteren können wiederum helfen, die Daten vorzustrukturieren und damit besser einzustufen. Dieses Vorstrukturieren hat man in Bilddaten ganz natürlich. Dort können wir sehr gut Strukturen erkennen, Linien, Geraden, Grauwertunterschiede. In genomischen Daten ist diese Struktur auch über eine chromosomale Lokation gegeben, aber nicht notwendigerweise und direkt über Funktionalität - wobei es Listen von Genen gibt, die eine funktionale Gruppierung erlauben, beispielsweise Onkogene - so dass man versucht, eine Struktur in den Daten im Einstufungsprozess zu verwenden. Ein weiterer Punkt wäre die Integration von Bilddaten. Da gibt es viele Methoden, die auch weiterentwickelt werden. Auf der algorithmischen Seite ist man hier nach meinem Dafürhalten schon weiter als dass es unbedingt groß angewandt werden würde.

Also müsste man erst die eingangs erwähnten Probleme lösen: Wie können die Daten zusammengeführt, ausgetauscht und geteilt werden?

Genau. Wie können sie anonymisiert und pseudonymisiert werden, damit man sie gut verwenden kann. Das Zugangsmanagement zu diesen Daten ist wahrscheinlich ein ganz entscheidender Punkt. Algorithmen werden von den Informatikern gern und manchmal auch schnell entwickelt. Man hat es mit Big Data zu tun, gerade bei genomischen Daten, aber das ist kein Vergleich zu Daten, die technischen Prozessen entstammen.

* Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/)

Fachbeitrag

26.03.2018
Walter Pytlik

Weitere Informationen

Prof. Dr. Hans A. Kestler
Leiter des Instituts für Medizinische Systembiologie
Universität Ulm
Tel.: +49 (0)731 50024500

- Medical Systems Biology

Der Fachbeitrag ist Teil folgender Dossiers

Data-Mining: Neue Chancen für Medizin und Gesundheit

Big Data – das große Versprechen der neuen digitalisierten Welt

Bioinformatik | Datenbank | Telemedizin | Data Mining | Big Data