

Data Mining in klinischen Datensätzen

Rasterfahndung

In der medizinischen Diagnostik zeichnet sich ein Paradigmenwechsel ab. Mit Data-Mining-Verfahren kann man aus ungezielt erhobenen Daten wertvolle Erkenntnisse gewinnen.

Unter Data Mining (deutsch „Datenschürfen“) fasst man eine Vielzahl von Computerverfahren zusammen, die alle dazu dienen, Erkenntnisse aus großen ungeordneten Datensätzen zu gewinnen. Ein berühmtes Beispiel ist die Videoüberwachung von U-Bahnhöfen: Der Computer analysiert kontinuierlich Millionen von Pixeldaten pro Sekunde und erkennt aus dem Muster automatisch, wenn sich eine kritische Situation abzeichnet.

Die Verlockung ist groß, solche mächtigen IT-Werkzeuge auch in der medizinischen Diagnostik einzusetzen. Vor allem unter dem Zeitdruck in den Aufnahme-stationen der Krankenhäuser, der durch immer kürzere Verweildauern erzeugt wird, zeichnet sich ein Paradigmenwechsel von der traditionellen Stufendiagnostik zur umfassenden Profilanalytik ab. Man möchte gern in wenigen Minuten möglichst viele Daten erheben und daraus die richtigen diagnostischen und organisatorischen Schlüsse ziehen.

Bereits vor über 30 Jahren versuchten Ärzte, aus ungezielten „Laborlatten“ mit Cluster- und Diskriminanzanalyse Diagnosen abzuleiten. Doch damals waren Laborcomputer zu leistungsschwach und die

Datensätze der „Autoanalyzer“ zu klein, so dass der Durchbruch ausblieb. Mit exponentiell steigender Computerpower und Zahl der Labortests gewinnt das Data Mining nun aber nach langer Vorlaufzeit auch in der klinischen Routine an Bedeutung, zum Beispiel in der Präventivmedizin und Krebsdiagnostik oder bei der Auswertung von Massenspektrometrie-, Flowzytometrie- und Biochipdaten.

Die Abbildung zeigt an einem einfachen Beispiel, was gemeint ist. Die Laborwerte aus der Aufnahmestation eines Akutkrankenhauses wurden völlig ungeordnet eingelesen und in standardisierte Farb-codes umgewandelt, um sie miteinander vergleichbar zu machen. Anschließend sortierte der Computer die Zeilen (Tests) und Spalten (Patienten) nach größtmöglicher Ähnlichkeit. Das Ergebnis sind so genannte Cluster (engl. Haufen), die gegenwärtiges medizinisches Wissen korrekt widerspiegeln.

So erkennt der Computer im linken Bildteil, dass Ery, Hb und Hk zusammengehören und dass vor allem solche Patienten erniedrigte Werte aufweisen, bei denen gleichzeitig die Harnstoff- und Kreatininwerte stark erhöht sind. Natürlich kennt der Computer den Grund für seine „Erkenntnis“ nicht: Die Blutbildung wird durch EPO aus der Niere angeregt; bei einem Nierenschaden versiegt die Produktion und es kommt zur renalen Anämie.

Dieses eher triviale Beispiel wurde ohne große Computerpower mit einem Excel-Programm erzeugt. Es soll nur zum Nachdenken über einen möglichen Paradigmenwechsel in der Labordiagnostik anregen.

Data Mining in großen klinischen Datensätzen ist eine echte Herausforderung. Für Diagnostica- und IT-Hersteller eröffnen sich hier neue Möglichkeiten, den Weg für eine schnelle und effiziente Diagnostik durch neue Technologien zu ebnet.

Buchbesprechung

Thomas A. Runkler
Data Mining – Methoden und Algorithmen intelligenter Datenanalyse
 65 Seiten, 72 Abb., 7 Tab.
 24,90 Euro
 ISBN 978-3-8348-0858-5
 Vieweg+Teubner Verlag.



Dieses Buch wendet sich an Leser, die mit Informationstechnologie bereits vertraut sind und Data Mining praktisch einsetzen möchten. Es bietet einen knapp gefassten und hervorragend strukturierten Überblick über die aktuellen Verfahren zur Extraktion von „Wissen“ aus großen Datensätzen. Geeignet für Naturwissenschaftler, Ärzte und Informatiker, für Einsteiger allerdings etwas schwere Kost.

Inhalt

Der Data-Mining-Prozess · Daten und Relationen · Datenvorverarbeitung · Visualisierung · Korrelation · Regression · Zeitreihenprognose · Klassifikation · Clustering.

Nutzanwendung für die Diagnostik¹

Das Data Mining in klinischen Datensätzen lässt sich vereinfachend in drei Schritte gliedern:

1. Datenaufbereitung
2. Exploration
3. Klassifizierung

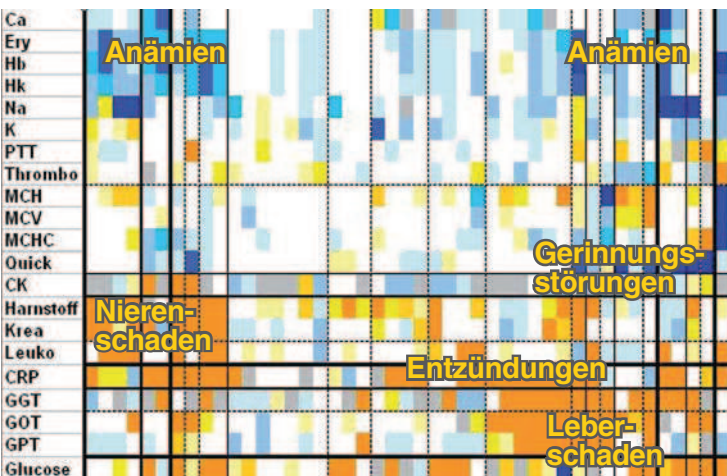
Bei einem typischen Data-Mining-Projekt verursacht die Datenaufbereitung etwa 80% des Gesamtaufwands, je 10% entfallen auf die Suche nach relevanten Wertemustern (Exploration) und deren Zuordnung zu bestimmten Krankheiten (Klassifizierung).

Kernpunkt der Datenaufbereitung ist die Normalisierung. Dabei werden die Absolutzahlen in Abhängigkeit von ihrem jeweiligen methoden- und gerätespezifischen Referenzbereich so umgerechnet, dass ein „normaler“ Wert zum Beispiel immer 0 ist (z-Transformation).

Die Exploration wird vor allem bei neuen Messverfahren eingesetzt, mit denen keine Erfahrungen vorliegen. Dies war zum Beispiel in den 1990er-Jahren bei Einführung der Biochips der Fall, als man noch nicht wusste, welche Krebsunterformen man damit entdecken würde.

Sind dann Testprofile und Diagnosen etabliert, kommen Klassifizierungsverfahren zum Einsatz. Sie gewichten und verdichten viele Messwerte zu einer einzigen Maßzahl (Klassifikator), die dann wie ein ganz normaler Laborwert über Referenzbereiche interpretiert wird.

¹ DGKL-Arbeitsgruppe Bioinformatik
www.dgkl.de/arbeitsgruppen/bioinformatik/



Das Ergebnis der explorativen Analyse ist eine „Datenlandkarte“, auf der sich interessante Farbreflekte – z. B. Verdachtsdiagnosen – aus dem Meer des Datenrauschens herausheben. In den Spalten stehen die Patienten, in den Zeilen die Tests (blau: erniedrigt, orange: erhöht, weiß: normal).

gh