

Data Mining in klinischen Datensätzen

Auszüge aus einem Bericht der DGKL-Arbeitsgruppe Bioinformatik

<http://www.dgkl.de/arbeitsgruppen/bioinformatik/>

Korrespondenzadresse
Prof. Dr. med. Georg Hoffmann
Trillium GmbH
Hauptstraße 12b
82284 Grafrath

Unter Data Mining (deutsch „Datenschürfen“) fasst man eine Vielzahl von Verfahren und Computeralgorithmen zusammen, die dazu dienen, neue Erkenntnisse aus großen, oft auch ungezielt erhobenen Datensätzen zu gewinnen. Erfolgreiche Anwendungen reichen von der Analyse des Käufer- oder Wählerverhaltens bis zur Rasterfahndung und Videoüberwachung.

Die Labordiagnostik beschäftigt sich seit über 30 Jahren mit Data Mining Techniken. Allerdings waren damals Laborcomputer zu leistungsschwach und die Datensätze aus den „Autoanalyzern“ der ersten Generation zu klein, um daraus neue Erkenntnisse zu gewinnen. Zwar wurde der höhere Informationsgehalt von Wertemustern im Vergleich zu Einzeltests bereits um 1990 belegt (1), doch dominierte das Paradigma der konsekutiven Abarbeitung (Stufendiagnostik, engl. *reflex testing*) dank leistungsfähiger Selektivanalysatoren der sog. zweiten und dritten Generation (2), so dass sich die Labordiagnostik nach dem Abklingen der ersten Begeisterung kaum noch mit Profilauswertungen auseinandersetzte.

Mit exponentiell steigender Zahl der messbaren Analyte und der Entwicklung einer neuen Generation hoch-paralleler Analysensysteme nimmt das Interesse nun aber wieder stark zu, zum Beispiel in der Genetik, Präventivmedizin oder Krebsdiagnostik, bei der Auswertung von Messprotokollen der Massenspektrometrie oder Durchflusszytometrie und bei der Evaluation von Microarrays und Mikrofluidikgeräten (3). Auch der DRG-bedingte Zeitdruck in den Krankenhäusern könnte einen Paradigmenwechsel von der traditionellen Stufendiagnostik zu einer schnellen, umfassenden Profildiagnostik, vor allem in der Aufnahme, begünstigen (4).

Die für das Data Mining benötigten Verfahren sind bekannt (5, 6). Einige werden von professionellen Anwendergruppen wie zum Beispiel der „R-Community“ (7) gepflegt, in der sich Biometriker und Bioinformatiker stark engagieren. Für Fragestellungen der Labormedizin eignen sich die dort bereitgestellten Algorithmen für die Analyse von Genexpressionsprofilen durchaus. Allerdings sind die meisten R-Programme für normale PC-Nutzer schwierig zu handhaben, berücksichtigen das spezifische Vorwissen der Labormedizin kaum und wirken auf den Nichtfachmann oft verwirrend.

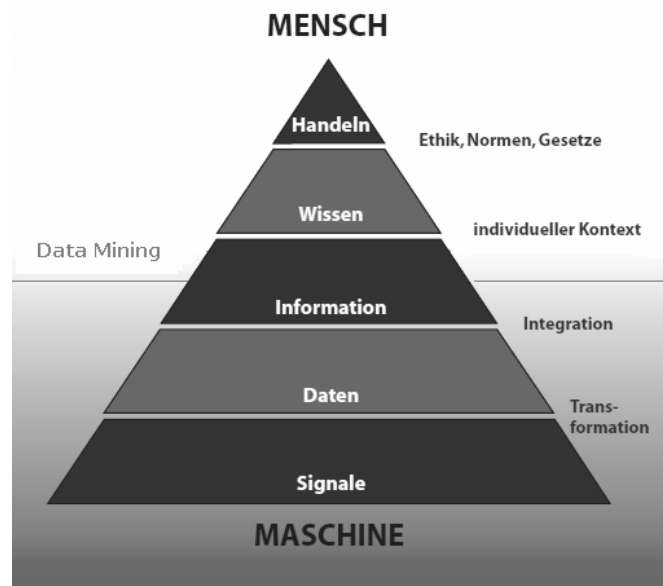
Data Mining ist eine logische Folge der Entwicklung von Datenbanktechniken, deren Anfänge in den 1950er-Jahren liegen. Ab etwa 1970 erlaubten relationale Datenbanken und die Abfragesprache SQL erstmals die gezielte Informationsverdichtung aus großen Datensätzen nach bestimmten Kriterien (`select ... from ... where ...`), und seit 1990 stehen fortgeschrittene Datenmodelle und Computersprachen sowie Vernetzungstechnologien zur Verfügung, um praktisch jede Information – ob Buchstaben, Zahlen, Bilder oder Töne – auf interessante Strukturen hin zu analysieren. Wer eine Melodie in sein Handy hinein singt und aus dem Internet den Titel des Liedes mit Komponist angezeigt bekommt, nutzt für diese einfache Data-Mining-Aufgabe bereits eine ziemlich hohe Komplexität von Datenbanken, Analyseverfahren und Schnittstellen.

Übertragen auf die medizinische Diagnostik müsste ein Data Mining Programm eigentlich in der Lage sein, aus allen zu einem Patienten vorliegenden Informationen – Anamnese,

Laborwerte, Röntgenbilder, Herztöne u.ä – Diagnose- und Therapievorschlage zu erstellen. Derartige Systeme gibt es wegen der im Vergleich zu einer Melodie viel hoheren Komplexitat eines Krankheitsbildes nicht und wird es moglicherweise auch nie geben, aber Teilschritte, fur die Computer besser geeignet sind als Menschen, lassen sich bereits gut definieren und wissenschaftlich untersuchen.

Der Fokus liegt dabei auf der *Datenaufbereitung*, *Exploration* und *Klassifizierung*. Es gibt zahlreiche weitere Data-Mining-Verfahren wie z.B. *online analytical* und *transactional processing* (OLAP, OLTP), Zeitreihen- und Bildverarbeitung, fur die hier auf die Lehrbucher verwiesen wird (5,6).

Zur Datenaufbereitung (*preprocessing*) gehoren z.B. Konsistenzprufung, Bereinigung und Zusammenfuhrung von Rohdaten sowie Kalibration, Qualitatsprufung, Filterung und Normalisierung von Messwerten. Diese scheinbar einfachen Schritte entscheiden letztlich uber Erfolg oder Misserfolg jedes Data Mining Projekts, benotigen viel domanen-spezifisches (also medizinisches) Fachwissen und verursachen in der Regel den groten Arbeitsaufwand. Deshalb grenzen einige Lehrbucher die Datenaufbereitung vom eigentlichen Data Mining ab, das domanen-unabhangige Techniken wie Datenbankabfragen, maschinelles Lernen oder statische Verfahren benutzt. Preprocessing und Data Mining werden dann unter einem eigenen Fachbegriff zusammengefasst, der immer wieder zu Missverstandnissen fuhrt: *knowledge discovery in data bases* (9). Medizinisches Wissen (*knowledge*) im engeren Sinne kann nur vom Menschen in einem individuellen Kontext generiert werden und ist die Voraussetzung fur arztliches Handeln. Der Computer hilft dabei lediglich, indem er Daten zu Information verdichtet. Data Mining unterstutzt den Menschen somit an der Schnittstelle zwischen Information und Wissen, ersetzt ihn aber nicht.



Vom Signal zum Handeln: Schritte arztlicher Informationsverarbeitung (modifiziert nach 10)

Beim Data Mining im Sinne dieses Berichts spielt der Begriff der *Klasse* eine zentrale Rolle. Sie wird definiert durch logische Eigenschaften, die alle ihre Objekte (Patienten, Tests) erfullen. So konnen Patienten einem bestimmten Krankheitsbild (Diabetes Typ I oder II), Stadium (fruh, spat) oder Organisationskriterium (ambulant, stationar) zugeordnet werden. Auch Tests lassen sich Krankheiten bzw. Organen (Tumormarker, Leberenzyme), Stadien (Fruherkennung, Verlaufskontrolle) oder Organisationskriterien (Aufnahmediagnostik, DRG-Kodierung) zuordnen.

Als Exploration bezeichnet man die Suche nach neuen, noch unbekanntem Klassen, als Klassifizierung die Vorhersage, in welche bereits bekannte Klasse ein neuer Tests oder Fall gehort. Die englischen Fachbegriffe dafur sind *class detection* und *class prediction*.

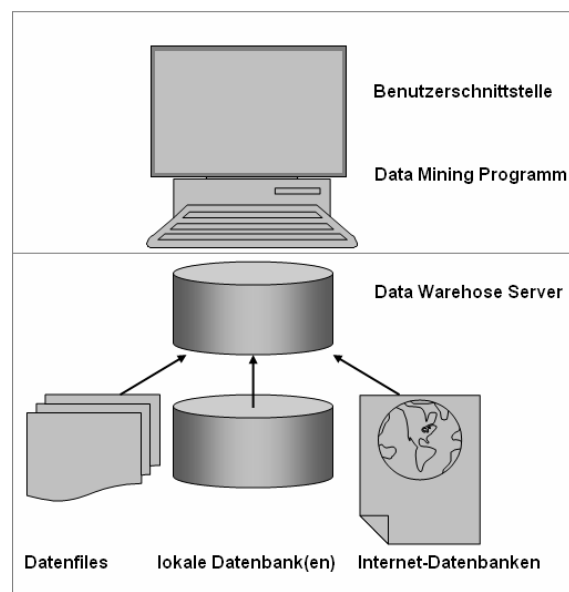
Data Mining mittels Exploration (*class detection*) ist in der Labordiagnostik vor allem dann sinnvoll, wenn man mit einer neuen Technologie experimentiert, über deren Aussagekraft man noch wenig weiß. So fanden Mosig et al (11) mit DNA-Microarrays insgesamt 2.318 Gene in Monozyten, deren Expression bei Patienten mit familiärer Hypercholesterinämie verändert war. Die Untersucher konnten sich dank ihres explorativen Ansatzes auf die wesentlichen 10% des Transkriptoms fokussieren und mit Hilfe von Datenbanken dann die relevanten Stoffwechselwege gezielt weiter untersuchen.

Im Gegensatz dazu operiert man bei Klassifizierungsprojekten (*class prediction*) mit deutlich weniger Daten. Haferlach et al (12) bestimmten zum Beispiel mit Microarrays in Leukozyten ein Set von 100 Genen, deren Expressionsprofil alle klinisch relevanten Leukämietypen mit 95 bis 100% Sensitivität und Spezifität unterscheiden konnte. Ein solcher multivariater Ansatz hilft also, mit einer einzigen neuen Technik ähnliche Ergebnisse zu erzielen wie mit einer Vielzahl traditioneller Verfahren von der Mikroskopie bis zur Chromosomenanalyse.

Die hier genannten Größenordnungen sind für das Data Mining in molekularbiologischen Datensätzen durchaus typisch: Aus einer Gesamtzahl von etwa 10^3 bis 10^6 verfügbaren Analyten (Gentranskripte, Proteine, Metabolite) werden durch Exploration einige hundert oder tausend „Kandidaten“ extrahiert, von denen dann zehn bis hundert in einen neuen „Klassifikator“ münden.

In klinischen Datensätzen liegen weit weniger Analyte vor, dafür ist die Heterogenität höher. Während man z.B. bei der Genexpressionsanalyse nur mRNA-Konzentrationen analysiert, enthalten klinische Befunde Konzentrationen verschiedener Substanzen, Enzymaktivitäten, Zellzahlen, Gerinnungszeiten u.v.m. – von halbquantitativen Laborergebnissen (++, <10) oder Blutdruckwerten (120/80) ganz abgesehen. Deshalb ist bei klinischen Daten die Aufbereitung für das Data Mining die eigentliche Herausforderung und nimmt in diesem Beitrag auch den größten Raum ein.

Bei professionellen Anwendungen ist das eigentliche Data Mining Programm je nach Größe des Projekts auf einem lokalen Rechner oder einem Server installiert und bezieht seine Informationen aus einer Datenbank oder einem Data Warehouse. Hier werden die zu verarbeitenden Laborwerte mit Zusatzinformationen wie Alter und Geschlecht, Referenzbereichen, Materialarten, Testnamen usw. verknüpft. Aus diesen Bausteinen setzt das Programm einen Bericht zusammen, der dann über die Benutzerschnittstelle präsentiert und gegebenenfalls weiter bearbeitet wird.



Typische Architektur eines Data Mining Systems: Der Benutzer sieht und bedient nur einen Ausschnitt aus dem Gesamtsystem, bestehend aus der Benutzerschnittstelle und dem Data Mining Programm. Dieses setzt auf einem Data Warehouse auf, das sich aus einer Vielzahl von Datenquellen (Studienprotokoll, KIS oder LIS, diverse Datenbanken) speist.

Die Aufbereitung und Zusammenführung der oft weit verstreuten Daten ist zeitaufwändig und erfordert an die Fragestellung angepasste IT-Werkzeuge bzw. gute Programmierkenntnisse. Im Idealfall existiert ein Programmassistent (engl. *Wizard*), der den Prozess im Dialog mit dem Benutzer abarbeitet, meist steht aber nur eine „Data Mining Engine“ für vorgegebene Inhalte und Formate zur Verfügung, die der Benutzer mit selbst aufbereiteten Daten befüllen muss. Für den Anfang eignet sich sicher MS Excel, aber bei großen Datenmengen ist die Aufbereitung mühsam und vor allem fehlerträchtig. Deshalb ist die Automatisierung häufig benötigter Aufgaben wie die Umrechnung von konventionellen in SI-Einheiten, die Auswahl oder Berechnung geeigneter Referenzbereiche und die Zuordnung lokal vergebener Namen zu internationalen Standards sinnvoll.

Normalisierung

Referenzintervalle und Entscheidungswerte repräsentieren wertvolles medizinisches Wissen. Für ihre standardisierte Ermittlung existieren ähnlich wie für die Testnomenklatur Vorschläge von amerikanischen (13) und europäischen (14) Organisationen. Gerade für neue Tests und innovative Testplattformen ist es aber nicht immer leicht, die geforderten Bedingungen zu erfüllen. Deshalb werden seit kurzem Schätzverfahren erprobt, die auf routinemäßig erhobenen Messwerten basieren (15, 16).

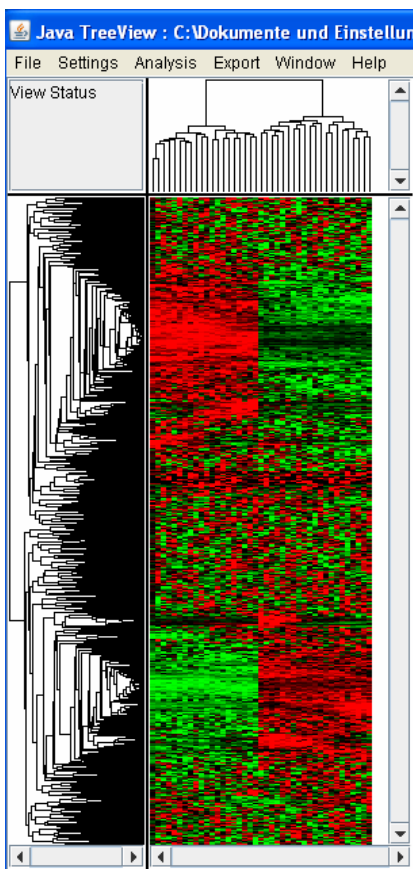
Das Haupteinsatzgebiet von Referenzwerten im Data Mining ist die Datennormalisierung. Sie dient in der Labordiagnostik und Molekularbiologie dazu, Messwerte vergleichbar zu machen, die auf unterschiedlichen Skalen und in verschiedenen Größenordnungen liegen. Im Idealfall streuen die Werte nach der Normalisierung nur noch in einem engen Bereich, beispielsweise von -10 bis +10. Ein erhöhter Wert von z.B. +5 sollte dann biologisch stets dasselbe bedeuten, gleichgültig ob der Originalwert mit Methode A oder B gemessen wurde, eine konventionelle oder SI-Einheit trägt, geringer oder hoher biologischer Variation unterliegt (z.B. Natrium versus CRP) usw.

Bisher erfüllt kein Normalisierungsverfahren diese Kriterien voll, aber für viele Data-Mining-Fragestellungen ist eine ungefähre Vergleichbarkeit ausreichend. In klinischen und methodischen Studien ist das *Autoscaling* (z-Transformation) am weitesten verbreitet. Es drückt die Abweichung jedes Messwerts vom Mittelwert des Kollektivs in Vielfachen der Standardabweichung aus. Das arithmetische Mittel der normalisierten Werte ist 0 und die Standardabweichung 1. Das Verfahren ist sehr einfach durchzuführen und entspricht der klassischen Definition der Normalisierung im Sinne einer Normalverteilung. Allerdings hat es in dieser einfachen Form zwei entscheidende Nachteile: Es ist nur auf symmetrisch verteilte Daten anwendbar und gewichtet pathologische Werte kaum stärker als normale.

Dieses Problem ist allerdings leicht zu lösen, wenn das Referenzintervall bekannt ist. In diesem Fall setzt man statt μ und σ den Mittelwert und die Standardabweichung des Referenzkollektivs μ' und σ' ein. Alle normalisierten Werte zwischen -2 und +2 sind fallen dann mit 95% Wahrscheinlichkeit in das Referenzintervall und ein Wert von +5 ist mit nahezu 100% Wahrscheinlichkeit erhöht.

Außerhalb der Labordiagnostik sind verteilungsunabhängige Verfahren sehr verbreitet, die anstelle der Standardabweichung Percentilen (100% = Gesamtbereich, 50% = Median usw.) einsetzen. In der Molekularbiologie und Bioinformatik wird schließlich die sehr einfache und stabile Log-Transformation, auch bekannt als „Pseudonormalisierung“, eingesetzt. Sie benötigt nur einen einzigen Referenzwert und liefert trotzdem ähnliche Werte wie das Autoscaling, wenn auch mit anderer Spreizung. Ein Messwert, der genau dem Referenzwert entspricht, ist auch hier 0 (denn $\log(1) = 0$), Werte von -1 bzw. +1 entsprechen der Hälfte bzw. dem Doppelten des Referenzwerts. Die Logarithmierung bewirkt, dass extreme Werteabweichungen nach oben und unten nivelliert werden, denn ein Wert von 2 entspricht einer Vervierfachung, von 3 einer Verachtfachung usw. So nähern sich die normalisierten Werte von Analyten wie Na und CRP aneinander an. Auch das Pareto, Range oder Power Scaling (34) verfolgen das Ziel, Extremwerte zu nivellieren. Ein Übersicht über die häufigsten Verfahren findet sich am Ende dieses Beitrags.

Auch wenn es auf den ersten Blick so aussieht, als sei die Normalisierung von Laborwerten eine Domäne von Data-Mining- und Bioinformatik-Experten, so hat sie doch auch in der Routine einen zunehmenden Stellenwert und wird als Lösung für aktuelle Probleme intensiv diskutiert: Normalisierung macht Laborwerte, die mit verschiedenen Methoden erhoben wurden, untereinander vergleichbar, erleichtert so den Methodenwechsel im Labor und beseitigt ganz nebenbei das Dilemma der Berichterstattung in konventionellen bzw. SI-Einheiten. Schließlich hilft sie, den hohen Informationswert quantitativer Daten zu erhalten, der vom Kliniker oft mangels Zeit oder Laborwissen auf die binäre Aussage „normal“ und „pathologisch“ reduziert wird. Wer nur auf das Sternchen schaut, trifft – vor allem bei schlecht definierten Referenzbereichen - häufig sogar die falsche Entscheidung (17). Stattdessen wird vorgeschlagen, Laborwerte mit abgestuften Farbtönen (rot für erhöht, grün für erniedrigt) zu unterlegen, wie es in der Bioinformatik seit langem üblich ist. So wird auf einen Blick klar, ob ein Wert normal, grenzwertig oder stark pathologisch ist.



In der Bioinformatik spielen Mustererkennung und die Datenfilterung eine große Rolle, um potenziell interessante Signale im Rauschen der irrelevanten Daten zu identifizieren. So ordnet die Clusteranalyse zusammengehörige über- oder unterexprimierte Gene (rot bzw. grün) so zusammen, dass man sie mit freiem Auge identifizieren und gezielt auswählen kann. Es gibt eine Vielzahl nützlicher lizenzfreier Programme wie zum Beispiel TreeView und GeneCluster. Sie benötigen stets speziell aufbereitete, insbesondere normalisierte Daten, um aussagekräftige Resultate zu liefern.

Weitere Schritte der Datenvorverarbeitung wie die Kalibration von Rohsignalen, Sicherung der Vergleichbarkeit von Experiment zu Experiment, Entfernung falscher Ergebnisse etc. spielen vor allem bei Microarray-Analysen eine große Rolle, sind aber in der Labordiagnostik durch Qualitätssicherungsprogramme gut abgedeckt und deshalb von geringerer Bedeutung. Bei sehr großen klinischen Datensätzen ist es jedoch sinnvoll, Tests wie Natrium und Chlorid herauszufiltern, die bei fast allen Patienten normal ausfallen und folglich – zumindest für das Data Mining – keinen Informationswert besitzen.

Ein großes und bisher ungelöstes Problem sind schließlich Wertelücken (*missing values*), da viele Data Mining Programme nur vollständige Datensätze bearbeiten können. Es gibt zwar zahlreiche Verfahren, um diese Lücken aufzufüllen, zum Beispiel mit dem Mittelwert des Gesamtkollektivs oder Werten von ähnlichen Patienten. In der Labordiagnostik ist dies aber in aller Regel nicht zulässig, denn selbst bei eng korrelierten Werten wie Harnstoff und Kreatinin kann einer der beiden normal, der andere pathologisch sein. Es ist besser, nur solche Data-Mining-Verfahren zu wählen, die mit Wertelücken umgehen können.

Literatur

1. Folkerts U, Nagel D, Vogt W. *The use of cluster analysis in clinical chemical diagnosis of liver diseases*. J Clin Chem Clin Biochem 1990; 28: 399-406
2. Hoffmann G. *Concepts for the third generation of laboratory systems*. Clin Chim Acta 1998; 278: 203-16
3. Hoffmann G. *Rasterfahndung*. Trillium Report 2009; 7:187
4. Hoffmann G. *Laborstrategien im Zeitalter von DRGs und Globalisierung*. Swiss Lab Med (Pipette) 2009; Heft 5: 12-15
5. Han J, Kamber M. *Data Mining, techniques and concepts*. Morgan Kaufmann Publishers, Elsevier, 2006
6. Runkler T. *Data Mining: Methoden und Algorithmen intelligenter Datenanalyse*. Vieweg und Teubner, 2009
7. Lehrstuhl für künstliche Intelligenz der Technischen Universität Dortmund: Internet-Programm *RapidMiner* (sourceforge.net/projects/yale/)
8. R Development Core Team: *Statistikpaket R* (www.r-project.org, ausführliche Beschreibung unter cran.r-project.org/doc/manuals/R-intro.pdf)
9. http://de.wikipedia.org/wiki/Knowledge_Discovery_in_Databases
10. Cullen P. Vom Signal zum Handeln. Trillium Report 2009; 7: 110-111
11. Mosig S, Rennert K, Büttner P, Krause S, Lütjohann D, Soufi M, Heller R, Funke H. *Monocytes of patients with familial hypercholesterolemia show alterations in cholesterol metabolism*. BMC Med Genomics 2008; 1:60. doi:10.1186/1755-8794-1-60
12. Haferlach T, Kohlmann A, Schnittger S, Dugas M, Hiddemann W, Kern W, Schoch C: *Global approach to the diagnosis of leukemia using gene expression profiling*. Blood 2005; 106:1189-98
13. *CLSI. Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline – third edition*. CLSI document C28-P3. Wayne,PA: Clinical and Laboratory Standards Institute; 2008;28:1-50.
14. Solberg H. *The IFCC recommendation on estimation of reference intervals. The RefVal program*. Clin Chem Lab Med 2004; 42: 710-4
15. Concordet D, Geffré A, Braun JP, Trumel C. *A new approach for the determination of reference intervals from hospital-based data*. Clin Chim Acta 2009; 405: 43-8
16. Arzideh F, Brandhorst G, Gurr E, Hinsch W, Hoff T, Roggenbuck L, Rothe G, Schumann G, Wolters B, Wosniok W, Haeckel R. *Ein verbesserter indirekter Ansatz zur Bestimmung von Referenzgrenzen mittels intra-laboratoriellen Datensätze am Beispiel von Elektrolyt-Konzentrationen*. J Lab Med 2009;33:52-66
17. Van den Berg R, Hoefsloot H, Westerhuis J, Smilde A, van der Werf M. *Centering, scaling, and transformations: improving the biological information content of metabolomics data*. BMC Genomics 2006; 7: 142, doi 10.1186/1471-2164-7-142
18. Sonntag O. *Über die Bedeutung und Interpretation des so genannten Normalwerts*. J Lab Med 2003; 27: 302-10

Name	Formel	Vorteile	Nachteile
z-Transformation* (Autoscaling) ^a	$\frac{x_i - \mu}{\sigma}$	Alle Tests werden gleich gewichtet	Empfindlich auf analytische Impräzision, unzulässig bei schiefen Verteilungen
Quantity quotient ^b	$10,2 \cdot \frac{x_i - \mu'}{\sigma'} + 100$	Intuitive Prozentwerte (analog IQ), bessere Gewichtung pathologischer Werte	Benötigt Referenzintervall, Probleme bei stark erniedrigten Werten
Range Scaling ^a	$\frac{x_i - \mu}{x_{\max} - x_{\min}}$	Wie z-Transformation, auch für schiefe Verteilungen geeignet	Sehr empfindlich auf Ausreißer
Dybkaer Normalisierung ^c	$\frac{x_i - m}{0,5 \cdot (Q_{0,68} - Q_{0,33})}$	Für schiefe Verteilungen geeignet, unempfindlich auf Ausreißer	Probleme bei geringer Streuung (0 im Nenner)
Log-Transformation (Pseudoscaling) ^d	$\log_2 \left(\frac{x_i}{R} \right)$	Sehr leicht zu berechnen, benötigt kein Streuungsmaß, gleicht Extremwerte aus	Nicht anwendbar auf Werte ≤ 0 , empfindlich auf sehr kleine und große Streuungen

Auswahl von Normalisierungsverfahren für das Data Mining in klinischen Datensätzen

x_i = Wert, μ = Mittelwert, σ = Standardabweichung, μ' = Mittelwert des Referenzintervalls*, σ' = Standardabweichung des Referenzintervalls*, m = Median, x_{\min} und x_{\max} = niedrigster und höchster Wert, $Q_{0,33}$ und $Q_{0,68}$ = Quantil 0,33 und 0,68 (sog. Perzentilen), \log_2 = dualer Logarithmus (zur Basis 2), R = Referenzwert

* Wenn das Referenzintervall mit den Grenzen R_{\min} bis R_{\max} bekannt ist, wird empfohlen, $\mu' = (R_{\min} + R_{\max})/2$ und $\sigma' = (R_{\max} - R_{\min}) / 4$ für die z-Transformation zu verwenden (unter der Annahme, dass das Referenzintervall symmetrisch ist und zwei Standardabweichungen nach oben und unten umfasst).