TAB
BÜRO FÜR TECHNIKFOLGEN-ABSCHÄTZUNG
BEIM DEUTSCHEN BUNDESTAG

Data-Mining – gesellschaftspolitische und rechtliche Herausforderungen

TAB-Fokus Nr. 40 zum Arbeitsbericht Nr. 203 Juli 2022

In Kürze

- Data-Mining steht für die Erkennung von Mustern und Strukturen in Datenbeständen. Generiert werden sowohl Informationen, z. B. zu Ähnlichkeiten, Abweichungen oder Auffälligkeiten, als auch mathematisch-statistische Modelle und Algorithmen, die in neuen Situationen des gleichen Sachverhalts eingesetzt werden können, um Entscheidungen zumindest zu unterstützen.
- Data-Mining werden große Innovationspotenziale in nahezu allen Lebensbereichen zugeschrieben.
- Bedenken gibt es wegen intransparenter Vorgehensweisen sowie ungleicher Verwertungsmöglichkeiten. Befürchtungen reichen bis zum Ende der Privatheit oder zur Unkontrollierbarkeit algorithmischer Systeme.
- Herausforderungen gibt es u. a. bei der Datenbereitstellung, der Konkretisierung der Möglichkeiten und Grenzen der Analyse sowie beim Umgang mit den Ergebnissen. Umfangreiche Fachkenntnisse sind für die Durchführung, Prüfung und Überwachung erforderlich.
- Folgeabschätzungen und Bewertungen sollten anwendungsbezogen erfolgen. Im Bericht werden Data-Mining-Beispiele in der Medizin und im Gesundheitssystem diskutiert.

Worum es geht

Data-Mining steht für die Anwendung mathematisch-statistischer Verfahren, um Strukturen und Muster in Datenbeständen zu erkennen. Wenn man damit verbundene gesellschaftliche Chancen und Herausforderungen erfassen will, sollte man nicht nur den unmittelbaren Einsatz solcher Verfahren (**Data-Mining im engeren Sinn**), sondern den gesamten Prozess der Informationsgewinnung aus Datenbeständen betrachten (**Data-Mining im weiteren Sinn**). Dazu gehören:

- **Aufgabendefinition**: Suche nach Ähnlichkeiten, Unterschieden oder Auffälligkeiten in Datenbeständen, Klassifikationen/Gruppierung von Objekten, Ableitung von Regeln, Modellierung
- **Datenauswahl und -aufbereitung**: Prüfung der Dateneignung, Fehlerbereinigungen, Umrechnungen, Erstellung von Analyse-/Trainingsdatensätzen
- **Datenanalyse**: Je nach Aufgabe und Datenform kommen unterschiedliche Verfahren in Betracht, einige gibt es seit Jahrzehnten (z. B. Cluster-/Regressionsanalysen), andere werden derzeit anwendungsreif (z. B. das Training künstlicher neuronaler Netze).
- **Ergebnisvalidierung**: Prüfung der ermittelten Parameter, Formeln oder Regeln

Data-Mining wird meist mit Analysen bereits vorhandener Datenbestände in Verbindung gebracht, die neu verknüpft oder zu neuen Zwecken weiterverwendet werden. Ein hoher Automatisierungsgrad ist beim unmittelbaren Einsatz der jeweiligen strukturerkennenden Verfahren auf einen standardisierten (Trainings-)Datensatz möglich, weil Algorithmen alle Analyseschritte definieren, die maschinell ausgeführt werden. Bei der Aufgabendefinition, der Datenaufbereitung und der Ergebnisprüfung sind vielfältige menschliche Tätigkeiten erforderlich. Meist werden in mehreren Schleifen (Zwischen-) Ergebnisse geprüft, das analytische Vorgehen angepasst sowie Daten hinzugezogen oder ausgeschlossen, bis die Resultate als ausreichend valide angesehen werden. Validitätsanforderungen variieren je nach Anwendungskontext. Erhebliches Sachverständnis ist nötig, um die Eignung der Daten und des mathematischen Analyseansatzes für die jeweilige Untersuchungsaufgabe einschätzen, Auffälligkeiten von Fehlern unterscheiden oder Ergebnisse interpretieren zu können.

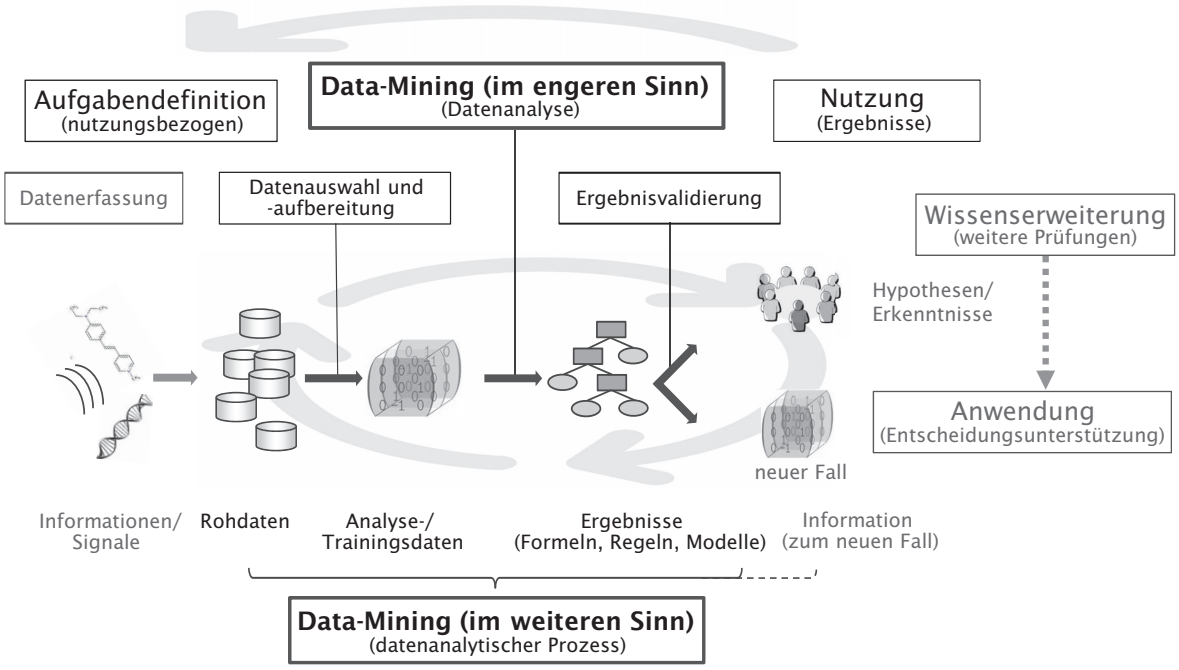
Das besondere Potenzial von Data-Mining besteht darin, dass ausreichend valide Ergebnisse generalisiert und weiter verwendet werden können, d.h. abgeleitete Entscheidungsregeln oder trainierte mathematisch-statistische Modelle können in neuen Situationen des gleichen Sachverhalts eingesetzt werden, um diese Situation zu bewerten und Entscheidungen zumindest

Auftraggeber

Ausschuss für Bildung, Forschung und Technikfolgenabschätzung
+49 30 227-32861
bildungundforschung@bundestag.de

ISSN-Print 2364-2610 | ISSN-Internet 2364-2629

Data-Mining: Darstellung der Prozessschritte



zu unterstützen. Derartige datenanalytische Vorgehensweisen sind seit jeher ein Kernelement wissenschaftlichen Arbeitens. Zunehmend werden auch in anderen Kontexten Informationsdienste bzw. algorithmische Entscheidungs(unterstützungs)systeme entwickelt und eingesetzt.

Data-Mining: Schlüsseltechnologie der Digitalisierung und gesellschaftliche Herausforderungen

Der Data-Mining-Begriff hat erhebliche Schnittmengen zu den Schlagworten Big Data, maschinelles Lernen oder Künstliche Intelligenz. Zahlreiche Gremien befassen sich seit einigen Jahren mit den Potenzialen und Herausforderungen der Digitalisierung im Allgemeinen sowie den wachsenden Datenbeständen, den Möglichkeiten und Grenzen der Weiterverwendung und Analyse sowie dem Umgang mit den Resultaten im Besonderen. Unisono wird empfohlen, Digitalisierungsaktivitäten zu forcieren, die Standardisierung und Normierung von Daten voranzutreiben, deren Bereitstellung über Dateninfrastrukturen zu verbessern, die Datennutzung stärker in den Blick zu nehmen, entsprechendes Know-how zu stärken, die Entwicklung datenanalytischer Anwendungen zu fördern, risikoreiche Anwendungen stärker zu regulieren sowie eine größere nationale oder europäische digitale Souveränität anzustreben, auch um hohe Schutzstandards und die Grundrechtesicherung im Umgang mit Daten zu gewährleisten. Diese Empfehlungen lassen sich auch aus den Ausführungen des TAB-Berichts ableiten.

Im TAB-Bericht werden nach der Darstellung technischer und rechtlicher Aspekte komplexer Datenanalysen im Allgemeinen einige öffentliche Aufgabenbereiche vertiefend betrachtet und bereichsspezifische Besonderheiten und Schwerpunktsetzungen herausgearbeitet. Dadurch können weitere bereichsbezogene Stärken und Handlungsoptionen abgeleitet werden.

Gute Data-Mining-Voraussetzungen im Geodatenbereich

Bei raumbezogenen Angaben haben sich Standards und Normen bereits bei der Datenerhebung, aber auch bei der Speiche-

Data-Mining-Anwendungsbeispiel in der Medizin

Die Mammografiebefundung wird seit Langem als ein möglicher Einsatzbereich für datentrainierte Algorithmen zur Objektklassifikation genannt, u. a. weil die durch nationale Screeningprogramme seit Jahren massenhaft erzeugten, fachärztlich befundeten und geprüften Mammografieaufnahme als Trainingsdaten genutzt werden könnten. Seit Jahren gibt es dazu vielfältiges datenanalytisches Engagement. Um solche resultierende Algorithmen regulär einsetzen zu

Radiologische Befundung von Mammografien

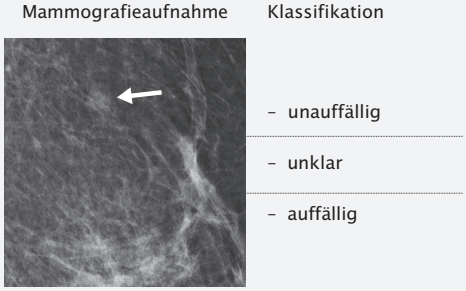


Abb. Kasten: nach Heindel et al. (2021): Systematische und qualitätsgesicherte Früherkennung des sporadischen Mammakarzinoms, S. 132 sowie nach Sickles et al. (2013): ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. 5th Edition. American College of Radiology, Reston

rung, Bereitstellung sowie bei einigen analytischen Funktionen weitgehend durchgesetzt. Seit Jahren wird die nationale Geodateninfrastruktur (GDI) ausgebaut, über die standardisierte amtliche Geodaten zunehmend offen bereitgestellt werden. Beim GDI-Ausbau gehört Deutschland zu den europäischen Vorreiterländern. Von den Erfahrungen könnten andere öffentliche Bereiche profitieren, die erst beginnen, Daten zu standardisieren und Dateninfrastrukturen aufzubauen.

Im Fokus der GDI-Aktivitäten lag bisher die Datenbereitstellung, nicht deren Weiterverwendung. Bisher lässt sich kaum abschätzen, wie intensiv bereitgestellte Geodaten für komplexe Datenanalysen genutzt werden. Perspektivisch sollte die Datennachfrage stärker in den Blick genommen werden. Das datenanalytische Potenzial von Geodaten steigt, je mehr Fachdaten georeferenziert bereitgestellt werden. Dazu sind erhebliche Anstrengungen in anderen Fachbereichen nötig.

Raumbezogene Analysen, vor allem wenn sie hochaufgelöste georeferenzierte Daten nutzen, können wichtige Informationen liefern (z. B. zur Lokalisierung gesundheitlicher Gefahren), aber auch Risiken der Grundrechteverletzung bergen (z. B. Tracking einzelner Personen) oder als sicherheitsbedenklich bewertet werden (z. B. Lageuntersuchungen von kritischen Infrastrukturen). Bisher gibt es eine gewisse Rechtsunsicherheit, wie bei raumbezogenen Analysen mit erhöhten Risiken vorzugehen ist. Die in Abstimmung befindliche europäische KI-Verordnung sollte diesbezüglich mehr Rechtssicherheit bringen.

Medizin: Erfahrungen mit Datentreuhandmodellen, Risikobewertungen und Qualitätssicherungen für datentrainierte Algorithmen

Medizinische Daten werden im Rahmen der Behandlung sowie vielfältiger Forschungs- und Entwicklungsaktivitäten in zunehmender Detailgenauigkeit erhoben. Die geringe Interoperabilität dieser dezentral gespeicherten, hochgradig geschützten, personenbezogenen medizinischen Primärdaten und der dafür erforderlichen IT-Systeme begrenzen deren Weiterverwendungsmöglichkeiten. Die Verbesserung der Interoperabilität medizinischer Primärdaten und IT ist eine vordringliche Aufgabe. Bisher sind aufwendige Datenaufbereitungen und Überführungen in spezielle medizinische Register und Datenzentren nötig, um sie weiterverwenden zu können. Nur ein Bruchteil der medizinischen Primärdaten wird derzeit in Repositorien überführt, die Datenbestände dauerhaft treuhändisch verwalten, kontinuierlich erweitern sowie bei Analyseanträgen Schutz- und Nutzungsinteressen im Einzelfall prüfen, bevor sie notwendige Daten einem begrenzten Analyseteam zugänglich machen. Deren Prüfverfahren sollten harmonisiert, vereinfacht und beschleunigt werden. Die Vernetzung der Repositorien und die Errichtung einer nationalen oder gar europäischen Gesundheitsdateninfrastruktur sind politische Ziele, deren Realisierung erheblicher Anstrengungen bedarf.

Werden anhand dieser Daten medizinische Entscheidungsregeln abgeleitet oder Modelle trainiert und diese zu behandlungsunterstützender Software weiterentwickelt, fällt diese Software unter das Medizinprodukterecht. Dieses Recht definiert risikoadjustierte Verfahren zur Qualitätssicherung der Produkte, die von Zertifizierungen anhand von Sicherheits- und Leistungsnachweisen bis zu kontinuierlicher Risikoüberwachung während der Anwendung reichen.

Sowohl die etablierten Treuhandstrukturen als auch die Verfahren zur Zertifizierung und Qualitätssicherung von Medizinprodukten können beispielgebend für andere risikoreiche Bereiche sein, in denen datenanalytische Verfahren und algorithmische Assistenzsysteme ebenfalls zunehmend eingesetzt werden (z. B Sicherheits-, Fin-Tech- oder Legal-Tech-Bereiche).

Auch mit umfangreichen Qualitätssicherungsmaßnahmen lassen sich beim Einsatz datentrainierter algorithmischer Systeme im Rahmen der Behandlung Risiken für Betroffene nie gänzlich ausschließen. Solche Systeme kommen bei sehr seltenen Situationen an ihre Grenzen, können real existierende Diskriminierungen reproduzieren und liefern Ergebnisse, die mitunter selbst für Expert/innen nur schwer nachvollziehbar sind. Deshalb ist die Klärung von dauerhaften Produktverantwortlichkeiten und von Haftungsfragen ein besonders relevanter Aspekt für die Akzeptanz und den Einsatz algorithmischer Assistenzsysteme, auch, aber nicht nur in der Medizin.

können, müssen zum einen deren Sicherheit und Leistungsfähigkeit geprüft werden (Medizinproduktezertifizierung). Diese Hürde nahmen bereits erste algorithmenbasierte Assistenzsysteme. Zum anderen müssen medizinische Fachgesellschaften und Kostenträger durch den Einsatz einen (Zusatz-)Nutzen zum nationalen Status quo anerkennen. Diese Hürde ist u. a. in europäischen Ländern, deren Screeningprogramme aufgrund der personalintensiven Mehrfachbefundung nur geringe Fehlerraten haben, besonders hoch. Deutsche Fachgesellschaften sind bezüglich algorithmenbasierter Assistenzsysteme zur Mammographiebefundung nach wie vor skeptisch. Sie halten die Umstellung von 2-D- auf 3-D-Aufnahmetechnologie für vielversprechender. Wenn sich die Aufnahmetechnologie ändert, müssen neue Trainingsdaten erstellt sowie Algorithmen erneut trainiert, geprüft und zertifiziert werden. In Ländern, in denen keine derart personalintensiven Befundroutinen etabliert sind, dürfte es leichter sein, mit derartigen Befundsystemen einen Nutzen zu erzielen.

Gesundheitssystemische Data-Mining-Projekte müssen einzeln bewertet werden

Eine hochrelevante Datenquelle für gesundheitssystemische Analysen sind die standardisierten Datensätze, mit denen medizinische Leistungen abgerechnet und vergütet werden. Sie bilden in der Summe das gesamte Leistungsgeschehen des ersten Gesundheitsmarktes auf Einzelfallebene ab und enthalten auch gesundheitsbezogenen Informationen zu allen Patient/innen, die in besonderem Maße geschützt sind. Die Weiterverwendung dieser Daten ist komplex reguliert. Zum einen dürfen diverse Institutionen diese Daten im Rahmen ihrer gesetzlich definierten Aufgaben nutzen. Zum anderen werden sie an zentrale Forschungsdatenzentren übermittelt. Es dauert jedoch oftmals Jahre, bis diese Versorgungsdaten in die Bestände der Zentren integriert sind und Dritte diese auf Antrag in sehr begrenztem Maße nutzen können. Eine zeitnahe Datenbereitstellung und eine breitere Nutzbarkeit werden vielfach gefordert.

Gesundheitssystemische Data-Mining-Ansätze sind oftmals eingebettet in komplexe Prozesse z. B. zur Leistungsvergütung und deren Fortschreibung, zur Qualitätssicherung der medizinischen Versorgung, zur Überwachung der gesundheitlichen Situation oder zur Suche nach unerwünschten Arzneimittelwirkungen. Inwiefern die jeweils genutzten Daten geeignet sowie die analytische Vorgehensweise zielführend und sinnvoll sind, ob dadurch medizinische Versorgungsprozesse verbessert und/oder gesundheitsbezogener Mehrwert generiert wird und welche unerwünschten Folgen damit einhergehen (z. B. gesundheitssystemische Fehlanreize), kann nur im Einzelfall bewertet werden.

Reichweite des Forschungsprivilegs für Data-Mining diskutieren

Datenanalysen sowie Data-Mining zu Forschungszwecken werden rechtlich auf unterschiedliche Art und Weise privilegiert. Einige Formulierungen sind jedoch auslegungswürdig. In der europäischen Datenschutz-Grundverordnung wird empfohlen, den Forschungsbegriff mit der Einhaltung anerkannter ethischer Forschungsstandards zu verknüpfen, ein entsprechendes Einwilligungsmanagement vorzusehen, Forschungsabsichten im Einzelfall zu prüfen und sowohl öffentliche als auch privatwirtschaftlich finanzierte Forschung bis hin zu technologischen Entwicklungen und Anwendungsdemonstrationen zuzulassen. Über Öffnungsklauseln werden jedoch nationale Spezifikationen zugelassen.

In der Medizin und im Gesundheitssystem sind ethische Forschungsstandards und Prüfungen von Analyseanträgen seit langem verankert. Das Einwilligungsmanagement ist jedoch eine Schwachstelle. Auch deshalb wird im nationalen Gesundheitssystem das Forschungsprivileg bisher enger ausgelegt und nur öffentlichen (Forschungs-)Einrichtungen, deren Analyseabsichten im öffentlichen Interesse liegen, ein Antrags- bzw. Nutzungsrecht gewährt. Parallel dazu sind Unternehmen, die klinische Studien zum Sicherheits- und Leistungsnachweis medizinischer Produkte finanzieren, nicht zur Bereitstellung ihrer Studiendaten zu Forschungszwecken verpflichtet.

Auf gesundheitsbezogene Risiken durch die Nichtnutzung existierender gesundheitsbezogener Daten wird seit Jahren hingewiesen, wenn z. B. Erkrankungsrisiken, Infektionsherde oder unerwünschte Nebenwirkungen von Behandlungsverfahren nicht erkannt werden oder Algorithmen anhand von Daten trainiert werden, die möglicherweise die nationale Situation nicht richtig repräsentieren. Diese Risiken durch Datennichtnutzung sollten in der Debatte um Datenschutz und -nutzung zukünftig stärker thematisiert werden.

TAB-Arbeitsbericht Nr. 203
Data-Mining – gesellschaftspolitische und rechtliche Herausforderungen
Katrin Gerlinger

Projektinformationen
www.tab-beim-bundestag.de/data-mining

Projektleitung und Kontakt
Dr. Katrin Gerlinger
+49 30 28491-108
gerlinger@tab-beim-bundestag.de

Das Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB) berät das Parlament und seine Ausschüsse seit 1990 in Fragen des wissenschaftlich-technischen Wandels. Das TAB ist eine organisatorische Einheit des Instituts für Technikfolgenabschätzung und Systemanalyse (ITAS) im Karlsruher Institut für Technologie (KIT). Zur Erfüllung seiner Aufgaben kooperiert es seit September 2013 mit dem IZT – Institut für Zukunftsstudien und Technologiebewertung gGmbH sowie der VDI/VDE Innovation + Technik GmbH. Der Ausschuss für Bildung, Forschung und Technikfolgenabschätzung entscheidet über das Arbeitsprogramm des TAB, das sich auch aus Themeninitiativen anderer Fachausschüsse ergibt. Die ständige »Berichterstattergruppe für TA« besteht aus dem Ausschussvorsitzenden Kai Gehring (Bündnis 90/Die Grünen) sowie je einem Mitglied der Fraktionen: Dr. Holger Becker (SPD), Lars Rohwer (CDU/CSU), Laura Kraft (Bündnis 90/Die Grünen), Prof. Dr. Stephan Seiter (FDP), Prof. Dr. Michael Kaufmann (AfD), Ralph Lenkert (Die Linke).