Data Mining – Ein neues Paradigma der angewandten Statistik

Marcus Hudec

Institut für Statistik und Decision Support Systems
Universität Wien

Gerhart Bruckmann zu seinem 70. Geburtstag gewidmet

Zusammenfassung: Im Rahmen dieses Beitrages wird versucht die Unterschiede im methodologischen Ansatz bei Verfahren des Data Mining den traditionellen statistischen Modellierungskonzepten gegenüberzustellen. Abschließend wird auf die allgemeinen Grenzen und Probleme bei der Anwendung von Methoden des Data Mining eingegangen.

Abstract: In this paper we try to compare differences in the methodological approach of Data Mining with traditional statistical concepts of modelling. Finally we will discuss general limitations and problems arising in the application of Data Mining algorithms.

Schlüsselwörter: Data Mining, Knowledge Discovery in Databases

1 Einleitung

Data Mining ist ein relativ junges in äußerst dynamischer Entwicklung stehendes Forschungsgebiet, welches sich im Schnittpunkt von verschiedenen Wissenschafts-disziplinen befindet. Konkret umfasst Data Mining ebenso Inhalte der Statistik wie Aspekte der Datenbanktheorie, des Machine Learning, des Pattern Recognition und der Artificial Intelligence - Forschung.

Zweifellos besteht in Bezug auf die Ziele und Aufgaben des Data Mining eine große Übereinstimmung mit jenen der klassischen angewandten Statistik und hier insbesondere der explorativen Datenanalyse. Für beide Disziplinen ist es eine primäre Zielsetzung relevante Strukturen und Muster in multivariaten Datenkörpern zu finden. Formulierungen wie „learning from data“ oder „turning data into information“ lassen sich gleichermaßen für beide Disziplinen anwenden.

Es wäre jedoch oberflächlich Data Mining als eine Subdisziplin der Statistik, als einen vorübergehenden Modetrend oder als eine wissenschaftliche Randdisziplin anzusehen. Vielmehr scheint die Empfehlung von Jon R. Kettenring (1997) angebracht, welcher in seiner Presidential Address vor der ASA (American Statistical Association) zum Thema Data Mining anmerkte: „*...we would be wise to pay very close attention and to become seriously involved with these developments*“.

In der gegenständlichen Arbeit vertreten wir daher die Ansicht, dass Data Mining als eine echte Bereicherung der Wissenschaftsdisziplin Statistik anzusehen ist. Folgt man Kuhns Theorie über Paradigmen in der Wissenschaftsgeschichte, so ist ein neues Paradigma dadurch charakterisiert, dass es dann von einer Forschergruppe als solches akzeptiert wird, wenn es einen wissenschaftlichen Kontext bietet, der es ermöglicht adäquate Antworten auf relevante aktuelle Fragestellungen zu geben.

Hiezu fand Tukey (1962) eine pointierte Formulierung: "*Far better an approximate answer to the right question, which is often wrong, than an exact answer to the wrong question, which can always be made precise.*"

Ein wesentliches weiteres Charakteristikum ist, dass ein neues Paradigma einer wissenschaftlichen Disziplin so weit gefasst sein muss, dass eine Vielzahl offener Fragen und Problemstellungen verbleiben, welche ein reiches Betätigungsfeld für neue Forschungsaufgaben eröffnen.

In diesem Sinne liefert Data Mining zum gegenwärtigen Zeitpunkt keineswegs endgültige Antworten. Gänzlich falsch wäre es aus einer Euphorie gegenüber dem Neuen altbewährte Denkmuster und Modelle der angewandten Statistik aufzugeben. Ebenso falsch wäre es aber, würde sich die statistische Community diesem neuen Paradigma verschließen und dieses spannende Forschungsgebiet anderen Wissenschaftsdisziplinen überlassen.

> "*Beware the Hype: The state of the art in automated methods in data mining is still in a fairly early stage of development. There are no established criteria for deciding which methods to use in which circumstances, and many of the approaches are based on crude heuristic approximations to avoid the expensive search required to find optimal, or even good solutions.*" (Fayyad, Piatetsky-Shapiro and Smyth, 1996)

2 Datenbanken als Ausgangspunkt wissenschaftlicher Analyse

Klassische Methoden der angewandten Statistik gehen von kleinen, überschaubaren Datensätzen aus, die häufig als unabhängige Realisierungen identisch verteilter Zufallsvariablen interpretiert werden können und oft spezifisch für die Beantwortung einer bestimmten Fragestellung gesammelt wurden. Datensätze mit einigen Tausend Beobachtungen gelten hier schon als groß.

Im Kontext des Data Mining sehen wir uns häufig mit großen Datenkörpern konfrontiert die Millionen von Datensätzen, welche keineswegs primär zum Zwecke der statistischen Analyse gesammelt wurden, umfassen. In diesem Sinne ist die Herausbildung von Data Mining eng mit der Computerisierung unserer Gesellschaft verknüpft. Die moderne Computertechnologie und Datenbanktechnik verfügt heute über kostengünstige Werkzeuge, die es uns ermöglichen riesige Datenmengen elektronisch zu speichern. Dabei ist es naheliegend, dass der Wunsch entsteht aus der reichen Vielfalt verfügbarerer Datensammlungen in Forschungsinstituten (z.B. Astronomie, Meteorologie) bzw. in Wirtschaftsunternehmen (z.B. Telekommunikationsunternehmen, Großbanken) nützliche Erkenntnisse zu gewinnen.

Der bekannte Trendforscher John Naisbett stellte einmal provokativ die These auf "*We are drowning in information, but starving for knowledge*".

Die Proponenten des Data Mining erkannten sehr bald, dass der klassische Approach der Statistik kein ausreichendes Werkzeug für diese neue Herausforderung liefert. "*To deal with the data glut, a new generation of intelligent tools for automated data mining and knowledge discovery is needed*". (Fayyad, Piatetsky-Shapiro, Smyth and Uthurusamy, 1996).

Vor diesem Hintergrund lassen sich die Verfahren des Data Mining summarisch als eine Methodensammlung zur Unterstützung des komplexen Prozesses der Identifikation valider, neuer, interessanter und letztlich auch nützlicher Muster und Regeln („Wissen") aus großen Datenbanken definieren. Häufig wird Data Mining auch daher mit *Knowledge discovery in databases* bezeichnet.

Abgesehen von der oft prohibitiven Größe der zu analysierenden Datenbasen, ist eine wesentlicher Unterschied im Entstehungsprozess der Daten auszumachen. Ausgangspunkt bilden Daten, die keineswegs primär zum Zweck statistischer Auswertungen erzeugt wurden, sondern vielfach als Ergebnis von operativen Transaktionen zu sehen sind. Die pragmatische Nutzung vorhandener Daten steht also im Vordergrund, während Fragen nach der Grundgesamtheit und theoretische Überlegungen über idealen Stichproben oder optimale Versuchspläne meist in den Hintergrund gedrängt werden. Dass daraus Probleme in Bezug auf die Qualität der Daten und die Interpretation der Ergebnisse resultieren ist offensichtlich und soll im Kapitel 4 diskutiert werden.

Die Größe der untersuchten Datenkörper sowohl in Bezug auf die Anzahl der Beobachtungen als auch in Bezug auf die Anzahl der Variablen stellen aber auch klassische Vorgehensmodelle der induktiven Statistik in Frage. In Anbetracht der großen Zahl an Beobachtungen erscheint z.B. das klassische Vorgehensmodell im Sinne von Neyman und Pearson, statistische Signifikanztests mit vorgegebenem Fehler 1.Art zu rechnen fragwürdig. Aufgrund der großen Fallzahl verfügen die Tests häufig über eine solche Power, dass selbst minimale, inhaltlich nicht relevante Effekte als statistisch signifikant eingestuft werden.

Auf der anderen Seite erscheinen Fragen wie multiples Testen und Aufrechterhaltung eines globalen Signifikanzniveaus beim Auswerten von Millionen von Assoziationen etwa bei der Auswertung von Einkaufsdaten einer Supermarktkette mit mehren tausend Artikeln nachrangig, da relative Aussagen in Bezug auf die Stärke der Assoziation von primären Interesse sind.

3 Unterschiede im methodologischen Ansatz

Methoden des Data Minings unterscheiden sich in verschiedenen Aspekten von Verfahren der klassischen Statistik.

Verfahren des Data Mining basieren typischerweise häufig auf problembezogenen „Ad-hoc"-Überlegungen. Im Gegensatz dazu basieren klassische statistische Verfahren auf stochastischen Modellen, die eine theoretische Analyse der Güte und Robustheit von Verfahren auf der Basis der mathematischen Statistik ermöglichen. Dieser Aspekt hat dazu geführt, dass Statistiker Methoden nach ihren Optimalitätseigenschaften unter

bestimmten theoretischen Konstellationen evaluieren und heuristischen Methoden, für welche keine allgemeinen theoretische Ergebnisse vorliegen, eher ablehnend gegenüberstehen.

> ”*An important feature of an estimator is consistency; in the limit, as the sample size increases without bound, estimates should almost certainly converge to the correct value of whatever is being estimated. Heuristic procedures, which abound in machine learning, have no guarantee of ever converging to the right answer*.” (Glymour et al., 1996).

Verfahren des Data Mining können demgemäß häufig nur in Bezug auf Ihre Tauglichkeit für bestimmte Aufgabenstellungen evaluiert werden. Aufgrund dieser Tatsache kommt vergleichenden Studien über die Leistungsfähigkeit von Algorithmen bzw. Methoden im praktischen Einsatz für konkrete Echtdaten eine wachsende Bedeutung zu. Im Kontext des Supervised Learning erscheint hier neben dem bereits klassischen StatLog-Projekt (Michie, Spiegelhalter und Taylor, 1994) in neuerer Zeit insbesondere die Arbeit von Lim und Loh (2000) erwähnenswert. In dieser Studie werden 34 Klassifikationsalgorithmen unterschiedlicher Provenienz in Bezug auf Klassifikationsgüte, Laufzeit und Komplexität anhand von 32 Datensätzen verglichen.

Bei den Methoden der klassischen Statistik steht nach wie vor das Modell im Vordergrund. In der Vergangenheit bediente man sich dabei relativ starrer nur gering parametrisierter Modelle, die zwar eine einfache Berechenbarkeit und theoretische Behandlung auch komplexer multivariater Fragestellungen gewährleisteten, sich jedoch oft als nicht robust gegenüber Verletzungen der Modellannahmen erwiesen und daher für die Praxis nicht immer adäquat waren. Die Entwicklung der multivariaten Statistik ist von einem Trend zu immer flexibleren komplexeren bzw. auch robusteren Modellen charakterisiert, welche jedoch selbst mit modernen Hochleistungscomputern insbesondere bei größeren Datensätzen beträchtliche computationale Probleme aufwerfen.

Beispielsweise sei hier auf die Entwicklung unterschiedlicher Verfahren des „model-based clustering“ (Banfield und Raftery, 1993) verwiesen, wo Clusterstrukturen unter verschiedenen Annahmen über die Mischung von multivariat normalverteilten Gruppen mittels Maximierung unterschiedlicher Kriterien gesucht werden. Dabei erweist sich die algorithmische Suche nach einer optimalen Partition bei großen Datenmengen als keineswegs trivial (Coleman et al., 1999; Hudec und Steiner, 2002).

Ein anderes Beispiel bei dem der Algorithmus die Eigenschaften des Schätzers quasi dominiert ist die robuste Schätzung von Kovarianzmatrizen (Woodruff und Rocke, 1994).

Die hier angesprochene Problematik trifft in noch verstärktem Maße auf viele Algorithmen des Data Mining zu. Der Verzicht auf globale Modelle, von dem diese Lösungsansätze typischerweise geleitet sind, rückt den Algorithmus und dessen computationale Umsetzung in den Vordergrund.

> “*The key role of programs has lead to an increased emphasis on algorithms in data mining, in contrast to the emphasis on models in statistics. The idea is that one applies the algorithm to data sets, learning how it behaves and what properties it has, regardless of any notion of an underlying model (or pattern) which it might be building.”* (Hand, 1999).

Ein weiterer Aspekt ist die Schwierigkeit Expertenwissen (sog. „Domain Knowledge“) im Rahmen der klassischen statistischen Methodologie in den Analyseprozess zu integrieren. Eine der wesentlichen Zielsetzungen des Data Mining ist es gerade Informationen aus unterschiedlichen Quellen möglichst in Form eines interaktiven Dialogs im Mining-Prozess zu berücksichtigen.

Als ein typisches Beispiel für Ansätze in diese Richtung wurde in der Machine Learning - Community schon sehr früh das sog. „conceptual clustering“ propagiert (Michalski und Stepp, 1983). Grundgedanke ist es dabei, dass ein Cluster-Algorithmus eventuell existierendes semantisches Vorwissen bei der Suche nach Clustern mitberücksichtigen können soll. Ein weiteres typisches Beispiel ist das Lernen von Klassifikationsregeln durch Attribut-orientierte Induktion, wie sie in einem der kommerziell erfolgreichsten Data Mining Pakete DB-Miner implementiert ist. Hier werden die einzelnen Tupel einer Relation durch eine Generalisierung von Attributwerten schrittweise in einer immer stärker vergröberten Relation zu einer Entität verschmolzen. Dabei werden semantische Hintergrundinformationen in Form von Abstraktionshierarchien für die vorkommenden Attributwerte genützt und entsprechende Klassifikationsregeln abgeleitet.

4 Allgemeine Probleme von Data Mining Ansätzen

4.1 Datenqualität

Data Mining darf nicht als eine losgelöste isolierte Aufgabe – die einmalige Analyse eines gegebenen Datensatzes – angesehen werden. Eine wesentliche Voraussetzung für die effiziente Nutzung von Mining Algorithmen insbesondere in der wirtschaftlichen Praxis ist deren Integration mit Data Warehouse Lösungen. Data Warehousing entspricht im Kontext des Data Mining dem Prozess der Stichproben- bzw. Erhebungsplanung in der klassischen Statistik, wodurch die Ausgangsbedingungen für die statistische Analyse definiert werden.

> ”*There is a symbiotic relationship between the activity of data mining and the data warehouse – the architectural foundation of decision support systems. The data warehouse sets the stage for effective data mining*.” (Inmon, 1996)

Einen wesentlichen Ansatzpunkt hierfür bieten die Konzepte der Metadaten-Modellierung, welche es erlauben Informationen über Inhalte und semantische Bedeutung der Daten eines Data Warehouse in Datenbankanwendungen zu integrieren und für Mining Zwecke verfügbar zu machen. Ein weiteres Charakteristikum einer guten Data Warehouse - Lösung ist, dass sie integrierte, fehlerbereinigte Daten auf unterschiedlichen Aggregationsniveaus in Form einer historischen Datenbank zur Verfügung stellt.

Es zeigt sich jedoch in der Praxis, dass Data Mining Algorithmen häufig unmittelbar auf Datenbanken von operativen Systemen angewendet werden, welche ungeprüfte und keineswegs konsolidierte Daten enthalten. Es lässt sich kaum vermuten, dass die unkritische Anwendung komplexer Algorithmen auf schlecht strukturierte Datenbanken zu reliablen und nützlichen Ergebnissen führen wird, weshalb die Gefahr besteht, dass

die derzeitige allgemeine Euphorie in der Geschäftswelt rund um das Thema Data Mining, welche zweifellos eine Chance für die Verbreitung statistischen Gedankenguts darstellt, schon bald mangels greifbarer Erfolge einer Ernüchterung weichen wird.

Typische Schwachstellen von in Datenbanken vorzufindenden Daten, welche die Anwendbarkeit von Data Mining Konzepten einschränken und auf die im Sinne einer statischen Qualitätssicherung hinzuweisen ist, sind:

- Mangelnde Repräsentativität (Fehlen relevanter Records)
- Selection Bias (Systematisches Fehlen von Records; ein typisches Beispiel ist die Problematik abgelehnter Kreditgewährungen im Credit Scoring)
- Mangelnde Charakterisierung durch Attribute (Fehlen wichtiger Variablen)
- Komplexe Korrelationsstrukturen (Confounding) bedingt durch fehlende Versuchsplanung
- Population Shift (Laufende Veränderung von Strukturen und Mustern in den datengenerierenden Prozessen)

4.2 Softwarequalität

Die gestiegene Verfügbarkeit komplexer Datenmengen in modernen Datenbanksystemen hat zu einem enormen Nachfragesog nach intelligenten Werkzeugen zur automatischen (oder zumindest weitgehend automatisierten) Analyse großer Datenmengen geführt. Die parallel dazu gestiegene Effizienz von Software – Entwicklungswerkzeugen und das massive Interesse großer Softwarehäuser bescherte uns in den letzten Jahren eine Flut von wissenschaftlichen Algorithmen sowie kommerziell verfügbaren Softwarepaketen.

Naturgemäß ergeben sich bei einer derart stürmischen Entwicklung Fragen nach der Qualitätssicherung und der Reliabilität der angebotenen Verfahren. Leidvolle Erfahrungen bei der historischen Entwicklung von klassischen statistischen Softwarepaketen sollten eine gewisse Grundskepsis gegenüber dem Einsatz ungeprüfter Analysesoftware bedingen.

Vor diesem Hintergrund kommen vergleichenden Evaluierungsstudien und wissenschaftlichem Benchmarking für kommerziell angebotene Data Mining Tools wachsende Bedeutung zu (Goebel und Gruenwald, 1999).

Ein häufig genanntes Charakteristikum von Methoden des Data Mining ist die Verwendbarkeit der Algorithmen für extrem große Datenmengen. Es zeigt sich jedoch in der Praxis, dass viele der kommerziell angebotenen Softwarepakete häufig durch die Größe des verfügbaren CPU-Memory beschränkt sind oder dass aus mangelnder Effizienz in der algorithmischen Umsetzung prohibitiv hohe Laufzeiten bei der Anwendung auf große Datenbasen resultieren.

4.3 Aussagekraft der Ergebnisse

Data Mining ist im wesentlichen ein exploratives Vorgehensmodell, bei dem Muster, Strukturen, Klassifikationsregeln und Hypothesen bzw. Erklärungsmodelle auf semi-automatische Weise direkt aus möglicherweise nicht repräsentativen Daten abgeleitet werden. In diesem Sinne handelt es sich also um eine hypothesengenerierende Vorgangsweise, deren Ergebnisse in Bezug auf die Anwendbarkeit auf andere Populationen bzw. Generalisierung nur mit größter Vorsicht interpretiert werden dürfen. Bei Data Mining geht es in der Regel nicht darum, „wahre Gesetzmäßigkeiten" über den datengenerierenden Prozess aufzuzeigen. Im Vordergrund steht, ob die Ergebnisse für den intendierten Zweck brauchbar bzw. praxistauglich sind.

Ein weiteres Problem kann im Überschätzen der Allmacht des Algorithmus liegen. Es besteht zweifellos die Gefahr, dass der Anwender jegliche Beziehung zu den Daten und ihrer Semantik verliert. Der komplexe Algorithmus wird für den Anwender zur undurchschaubaren Black-Box, die ihm von den Daten trennt. Das Überprüfen von Modellannahmen entfällt, und die vom Algorithmus generierten Ergebnisse, welche oft nur eines von vielen mögliche Interpretationsszenarien eines komplexen Datensatzes darstellen, werden fälschlicherweise als erwiesenes Faktum angesehen.

Häufig erlauben empirischen Daten keine eindeutige Entscheidung zwischen in Bezug auf die den Algorithmus steuernden Kriterien nahezu äquivalenten Modellen, welche jedoch eine unterschiedliche semantische Interpretation aufweisen. Sensitivitäts-analysen und Visualisierungstechniken werden hier zum unverzichtbaren Instrument, will man das Auffinden von Artefakten vermeiden.

In diesem Kontext erscheint ein Zitat von Pregibon beachtenswert, der darauf hinweist, dass es der statistischen Community offensichtlich nicht gelingt die Vorteile klassischer Methoden aufzuzeigen und einen angemessenen Stellenwert innerhalb des Data Mining einzunehmen.

> „*The tendency of the statistical community to propagate uncertainty in their models through sampling distributions, their familiarity with the need to regularize models, and their dogged perseverance in checking models assumptions and stability are strengths.*
>
> *Still, alternative heuristic modelling techniques have gained in popularity partly as a way to avoid statistics, yet still address challenging induction tasks.*
>
> *Statisticians should learn from this need to do a better job of communicating the value of such considerations, as well as clarifying and streamlining ways of injecting extra-data information into the modelling process.*"

Dabei wird die Relevanz klassischer statistischer Konzepte selbst von den führenden Proponenten des Data Mining außer Zweifel gestellt:

> „*Data Mining carried out poorly (without regard to statistical aspects of the problem) is to be avoided*. " (Fayyad, Piatetsky-Shapiro and Smyth, 1996).

4.4 Privacy und Datenschutz

Mit der wachsenden Verfügbarkeit von elektronisch gespeicherten Daten eröffnen sich nicht nur immer mehr vielversprechende Anwendungsmöglichkeiten des Data Mining, gleichzeitig wächst auch die Gefahr einer missbräuchlichen Verwendung von Datenbasen.

Ein ethisch verantwortungsvoller Umgang ist hier, ebenso wie die Entwicklung von softwaretechnischen Schranken zur Vermeidung unerwünschter Datenverknüpfungen, auf der Ebene von Individualdaten gefordert.

Naturgemäß ergeben sich hier auch Herausforderungen für die Legislative, um geeignete Rahmenbedingungen zu schaffen.

5 Epilog

Viele der vom Data Mining fokussierten Probleme sind keineswegs neu. Der Umstand dass Data Mining erst in den letzten Jahren als eigenständiges Forschungsgebiet aufblühte, ist eng mit der technologischen Entwicklung auf dem Computersektor verknüpft.

Betrachtet man die Tätigkeit von Gerhart Bruckmann, als „Wahlhochrechner der Nation“, so wird deutlich, dass der Jubilar in mehrerer Hinsicht aus dem engen Korsett des klassischen Paradigmas der Angewandten Statistik ausgebrochen ist.

Alleine die Wahl der Vorgangsweise und der Aufgabenstellung ist charakteristisch für die Methodologie des modernen Data Mining. Im Gegensatz zu klassischen Stichprobenverfahren zur Prognose von Wahlergebnissen, welche aufgrund des gewählten Stichprobenverfahrens einen exakten Stichprobenfehler ermitteln können, verwendete Bruckmann bei seinen Wahlhochrechnungen einfach pragmatisch unmittelbar jene Daten, die im Zuge des Auszählprozesses anfielen. Naturgemäß erschwert dies die exakte Interpretation von Konfidenzintervallen für Schätzungen im Sinne der mathematischen Statistik.

Ein weiterer Aspekt ist der Umfang der Datenbasis, welcher in Relation zur damals verfügbaren Computerleistung zu sehen ist. Erschwert durch den Umstand der Aktualität und des damit verbundenen Zeitdrucks kamen auch hier nur solche Algorithmen für Hochrechnungszwecke in Frage, welche eine automatisierte Analyse ermöglichten, ohne dass der Statistiker unmittelbar über Einzeldaten reflektieren konnte. Dabei setzte Bruckmann nicht einen im Sinne eines theoretischen Kriteriums optimalen Algorithmus ein, sondern verwendete ein Spektrum unterschiedlicher auf verschiedenen Modellannahmen beruhender Hochrechnungskonzepte, um auf diese Weise durch unterschiedliche Szenarien zusätzliche Information über die Schwankungsbreite der Hochrechnung zu erhalten.

Nicht zuletzt sei aber auch erwähnt, dass Bruckmann die Relevanz externer über die Daten hinausgehender Information erkannt hatte und entsprechend berücksichtigte. So parametrisierte Bruckmann bei einer seiner erfolgreichsten Hochrechnungen - der korrekten Prognose der knappen Atomvolksabstimmung – sein Modell dergestalt, dass er die Ergebnisse des Bundeslandes Vorarlberg nicht für die Trendhochrechnungen der anderen Bundesländer berücksichtigte. Diese Entscheidung war insofern gravierend, da

die Ergebnisse Vorarlbergs traditionell bereits frühzeitig verfügbar waren und Bruckmann somit auf eine Fülle von Ergebnissen bei den Trendhochrechnungen der anderen Bundesländer „verzichtete“. Die Nutzung der semantischen Hintergrundinformation, dass sich Vorarlberg bei diesem Votum extrem abweichend verhalten würde, über welche Bruckmann als Kenner der aktuellen sozio-politischen Landschaft Österreichs verfügte, bildete die Basis für die korrekte Hochrechnung.

Es wäre wohl im Überschwang einer Festschrift zu weit gegriffen, würde man Bruckmann nun als einen der ersten „Data Miner“ bezeichnen. Parallelitäten lassen sich aber, wie oben skizziert, unschwer erkennen. Charakteristisch für das Schaffen des Jubilars ist sein Bestreben, relevante empirische Fragestellungen durch den innovativen, die Grenzen hergebrachter Denkmuster sprengenden Einsatz formaler Methoden unter Einbeziehung aller verfügbaren Informationen zu beantworten.

Literatur

J.D. Banfield and A.E. Raftery. Model based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803-821, 1993.

D. Coleman, X. Dong, J. Hardin, D.M. Rocke, and D.L. Woodruff. Some computational issues in cluster analysis with no a priori metric. *Comp. Stat. and Data Analysis*, 31:1-11, 1999.

U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, 1996.

U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 37-54, 1996.

C. Glymour, D. Madigan, D. Pregibon, and P. Smyth. Statistical inference and data mining. *CACM*, 39:35-41, 1996.

M. Goebel and L. Gruenwald. A survey of data mining and knowledge discovery software tools. *SIGKDD Explorations*, 1:20-33, 1999.

D.J. Hand. Why data mining is more than statistics writ large. In *Bulletin of the International Statistical Institute ISI 99*, Proceedings 52nd Session, pages 433-436, 1999.

M. Hudec and P.M. Steiner. Model-based classification of large data sets. To appear COMPSTAT 2002.

W.H. Inmon. The data warehouse and data mining. *CACM*, 39: 49-50, 1996.

J.R. Kettenring. Shaping statistics for success in the 21st century. *JASA*, 92:1229-1234, 1997.

T. Lim and W. Loh. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning*, 40(3):203-228, 2000

R.S. Michalski and R.E. Stepp. Learning from observation; Conceptual clustering. In M. Kaufmann, editor, *Machine Learning: An Artificial Intelligence Approach*. pages 331-363, 1983.

D. Michie, D.J. Spiegelhalter, and C.C. Taylor. *Machine Learning, Neural and Statistical Classification*. Englewood Cliffs, N.J, 1994.

J.W. Tukey. The future of data analysis. *Annals of Mathematical Statistics* 33:1-67, 1962

D.L: Woodruff and D.M. Rocke. Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *JASA*, 89:888-896, 1994.

Adresse des Autors:

Ao.Univ.Prof. Dr. Marcus Hudec
Institut für Statistik und Decision Support Systems
Universität Wien
Universitätsstraße 5/3
A-1010 Vienna
Austria

Tel. +43 1 4277 / 38622
Fax +43 1 4277 / 9386
E-Mail: marcus.hudec@univie.ac.at
http://mailbox.univie.ac.at/Marcus.Hudec/