Data Mining mit Neuro-Fuzzy-Systemen

Rudolf Kruse, Christian Borgelt und Detlef Nauck
Institut für Wissens- und Sprachverarbeitung
Otto-von-Guericke-Universität Magdeburg

Zusammenfassung: Unter Data Mining versteht man die Anwendung von Modellierungs- und Entdeckungstechniken, um neue, unerwartete, valide, verständliche und verwertbare Informationen aus Datenbanken zu gewinnen. Zum Einsatz gelangen neben Methoden aus dem Datenbankbereich (Data Warehousing) und der Statistik (explorative Datenanalyse) auch Nicht-Standardansätze aus verschiedensten Bereichen wie z.B. der Neuronalen Netze, des Maschinellen Lernens (z.B. Entscheidungsbäume und Induktive Logische Programmierung) und der Fuzzy-Systeme, wobei letztere besonders wegen ihrer Einfachheit bei Anwendern sehr beliebt sind. In diesem Aufsatz konzentrieren wir uns auf Methoden aus dem Bereich der Neuro-Fuzzy-Systeme, die in Kooperation mit verschiedenen Industriefirmen entwickelt und erfolgreich eingesetzt wurden.

1 Einleitung

Durch die moderne Informationstechnologie, die jedes Jahr leistungsfähigere Rechner hervorbringt, ist es heute möglich, mit sehr geringen Kosten große Datenmengen zu sammeln und zu speichern. Folglich kann es sich eine immer größer werdende Zahl von Unternehmen, wissenschaftlichen Institutionen und Behörden leisten, große Archive von Daten, Zahlen, Texten, Bildern etc. aufzubauen. Es zeigt sich jedoch, daß es sehr schwer ist, das in diesen Datenbeständen verborgene Wissen auszunutzen. Im Gegensatz zu der Flut an Daten gibt es einen Mangel an Werkzeugen und Methoden, mit denen sich aus Daten die nützlichen Informationen und Muster gewinnen lassen. Obwohl ein Anwender gewöhnlich ein ungefähres Verständnis seiner Daten und ihrer Bedeutung hat — so kann er etwa Hypothesen formulieren und Abhängigkeiten erraten —, weiß er doch selten wo die "interessanten" und "relevanten" Informationen zu finden sind, ob diese Informationen seine Hypothesen und Modelle stützen, ob (andere) interessante Phänomene in den Daten verborgen sind, welche Methoden am besten geeignet sind, um die benötigten Informationen schnell und verläßlich zu finden, und wie die Daten in menschliche Begriffe übersetzt werden können, die für den betreffenden Anwendungsbereich angemessen sind.

Als Antwort auf diese Herausforderungen hat sich ein neues Forschungsgebiet entwickelt, das „Knowledge Discovery in Databases“ (KDD) oder „Data Mining“ (DM) genannt wird. Es wird üblicherweise wie folgt charakterisiert [8]:

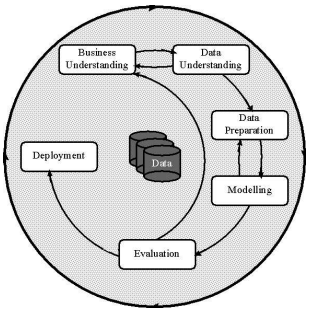
> Knowledge discovery in databases (KDD) is a research area that considers the analysis of large databases in order to identify valid, useful, meaningful, unknown, and unexpected relationships.

Oft wird der Begriff „Data Mining“ auf den Schritt des Wissensentdeckungsprozesses beschränkt, in dem Entdeckungs- und Modellierungstechniken angewendet werden. Data Mining wird dann definiert durch eine Liste von Aufgaben [8, 15], die wenigstens die folgenden umfaßt: Segmentierung (z.B.: Welche Arten von Kunden hat ein Unternehmen?), Klassifikation (z.B.: Ist diese Person ein potentieller Kunde?), Konzeptbeschreibung (z.B.: Welche Eigenschaften hat ein potentieller Kunde?), Voraussage (z.B.: Wie wird sich der Dollarkurs entwickeln?), Abweichungsanalyse (z.B.: Gibt es saisonale Umsatzschwankungen und was sind ihre Gründe?) und Abhängigkeitsanalyse (z.B.: Welche Produkte werden in einem Supermarkt häufig zusammen gekauft?).

Data-Mining-Methoden kommen aus sehr verschiedenen Bereichen, wie z.B. Statistik, Soft Computing, Künstliche Intelligenz und Maschinelles Lernen. Bekannt sind z.B. Regressionsanalyse, Diskriminanzanalyse, Zeitreihenanalyse, Entscheidungsbäume, (Fuzzy-)Clusteranalyse, Neuronale Netze, Induktive Logische Programmierung, Assoziationsregeln usw. (Ein Vergleich einiger kommerzieller Softwareprodukte, die diese Methoden anbieten, findet sich z.B. in [10].) In diesem Aufsatz konzentrieren wir uns auf Neuro-Fuzzy-Methoden und zeigen wo und wie sie eingesetzt werden können.

Mit dem Einsatz von Fuzzy-Methoden betont man die Forderung nach *verständlichen*, am besten in sprachliche Regeln gefaßten Ergebnissen, denn die *Fuzzy-Mengen-Theorie* ist ein ausgezeichnetes Mittel, um die „unscharfen“ („fuzzy“) Grenzen sprachlicher Ausdrücke zu modellieren. Im Gegensatz zur klassischen Mengenlehre, in der ein Objekt oder Fall entweder Element einer (z.B. durch eine Eigenschaft definierten) Menge ist oder nicht, erlaubt die Fuzzy-Mengen-Theorie, daß ein Objekt oder Fall nur zu einem gewissen Grade zu einer Menge gehört. Sie modelliert so die „Penumbra“ des Begriffs, der die definierende Eigenschaft bezeichnet [12]. Zugehörigkeitsgrade können als *Ähnlichkeit*, *Präferenz* oder *Unsicherheit* gedeutet werden [6]: Sie können angeben, wie ähnlich ein gegebenes Objekt oder ein vorliegender Fall zu einem prototypischen ist, sie können Präferenzen zwischen suboptimalen Problemlösungen darstellen oder sie können Unsicherheit über die wahre Situation ausdrücken, wenn diese durch unscharfe Begriffe beschrieben ist. Wegen ihrer Nähe zum menschlichen Denken sind Fuzzy-Lösungen gewöhnlich leicht zu verstehen und anzuwenden. Sie sind daher bevorzugte Methoden, wenn sprachliche, vage oder impräzise Informationen verarbeitet werden müssen [14].

2 Fuzzy-Methoden im Data Mining



Das CRISP-DM-Modell

Die Forschung im Bereich der Wissensentdeckung in Datenbanken und des Data Mining hat zu einer großen Zahl von Vorschlägen für ein allgemeines Modell des KDD-Prozesses geführt. Ein neuerer Vorschlag, der vermutlich sehr einflußreich sein wird, da er von mehreren großen Firmen wie NCR und DaimlerChrysler unterstützt wird, ist das CRISP-DM-Modell (CRoss Industry Standard Process for Data Mining) [5]. Die Grundstruktur dieses Prozeßmodells ist rechts gezeigt. Der Kreis deutet an, daß es sich im wesentlichen um einen zirkulären Prozeß handelt, in dem die Bewertung der Ergebnisse ein erneutes Ausführen der Datenauswahl und -aufbereitung und der Modellbildung anstoßen kann. In diesem Prozeß können Fuzzy-Methoden in mehreren Phasen nutzbringend eingesetzt werden:

Die Phasen des *Anwendungsverstehens (business understanding)* des *Datenverstehens (data understanding)* sind i.a. stark auf menschliche Leistungen angewiesen und lassen sich kaum automatisieren. In diesen Phasen werden i.w. die Ziele des Wissensentdeckungsprojektes definiert, der potentielle Nutzen eingeschätzt, und die notwendigen Daten identifiziert und zusammengeführt. Außerdem wird Hintergrundwissen und Meta-Wissen über die verfügbaren Daten gesammelt. In diesen Phasen können Fuzzy-Methoden benutzt werden, um z.B. vage formuliertes Hintergrundwissen so zu repräsentieren, daß es in der späteren, stark automatisierbaren Modellierungsphase genutzt werden kann. Weiter sind Fuzzy-Datenbankanfragen nützlich, um die benötigten Daten zu finden und um zu prüfen, ob es sinnvoll ist, zusätzliche, verwandte Daten heranzuziehen.

In der *Datenaufbereitung (data preparation)* werden die gesammelten Daten von Fehlern und Ausreißern gereinigt, in ein passendes Format überführt und ggf. geeignet skaliert, um Eingabedatensätze für die Modellierungstechniken zu erhalten. In diesem Schritt können Fuzzy-Methoden z.B. genutzt werden, um Ausreißer zu erkennen, etwa indem man ein *Fuzzy-Clustering*-Verfahren [3, 11] anwendet und anschließend diejenigen Datenpunkte bestimmt, die weit entfernt von den Clusterzentren liegen.

Die *Modellierungsphase (modeling)*, in der Modelle gebildet und an die Daten angepaßt werden, um z.B. die zukünftige Entwicklung vorherzusagen oder um Klassifikatoren zu erhalten, kann offenbar am stärksten vom Einsatz von Fuzzy-Methoden profitieren. Man kann i.w. zwei Klassen von Fuzzy-Methoden unterscheiden: Die erste Klasse, die *Fuzzydaten*-Analyse [13], besteht aus Verfahren zur Analyse von unscharfen Daten, d.h. Daten, die von ungenauen Meßinstrumenten geliefert werden oder sich aus vagen Aussagen menschlicher Experten ergeben. Ein Beispiel aus unserer eigenen Forschung ist die Induktion von *possibilistischen graphischen Modellen* [4] aus Daten, die der bekannten Induktion probabilistischer graphischer Modelle analog ist. Die zweite Klasse, die Fuzzy-*Datenanalyse* [1], besteht aus Verfahren, die Fuzzy-Methoden einsetzen, um scharfe Daten zu strukturieren und zu analysieren, z.B. *Fuzzy-Clusteranalyse* zur Datensegmentierung und Regelerzeugung und *Neuro-Fuzzy-Systeme* zur Regelerzeugung.

In der *Bewertungsphase (evaluation)*, in der die Ergebnisse getestet werden und ihre Qualität bestimmt wird, zeigt sich die Nützlichkeit von Fuzzy-Modellierungstechniken am deutlichsten. Da sie zu verstehbaren Systemen führen, können sie leicht auf Plausibilität geprüft und mit der Intuition und den Erwartungen eines menschlichen Experten verglichen werden. Außerdem liefern sie, im Gegensatz zu Neuronalen Netzen, die „Black Boxes“ sind, oft wichtige neue Einsichten in den Anwendungsbereich.

3 Regelerzeugung mit Neuro-Fuzzy-Systemen

Als Beispiel für die Verwendung von Fuzzy-Methoden im Data Mining betrachten wir Neuro-Fuzzy-Systeme zur Regelerzeugung. Ein Fuzzy-System besteht gewöhnlich aus einer Regelbasis (Struktur) und Fuzzy-Partitionen der Wertebereiche aller Variablen (Parameter). Diese Bestandteile gilt es durch eine Analyse der verfügbaren Daten zu bestimmen, wobei zu beachten ist, daß es oft einen Antagonismus zwischen Genauigkeit und Interpretierbarkeit gibt. Gerade Fuzzy-Systeme werden nicht nur anhand der Exaktheit der mit ihnen erzielbaren Ergebnisse, sondern auch, wenn nicht sogar vornehmlich, anhand ihrer Einfachheit und Verständlichkeit beurteilt. Gewöhnlich werden sie gerade dann verwendet, wenn der Anwender durch Untersuchung einer gelernten Regelbasis Einsichten in die Zusammenhänge in seinen Daten gewinnen will. Wichtig für die Verständlichkeit

eines Fuzzy-Systems sind vor allem, daß

- die Zahl der Regeln in der Regelbasis klein ist,
- in einer Regel nur wenige Variablen verwendet werden,
- die Wertebereiche durch wenige, sinnvolle Fuzzy-Mengen partitioniert sind, und
- kein linguistischer Ausdruck durch mehr als eine Fuzzy-Menge dargestellt wird.

Es gibt verschiedene Möglichkeiten, ein Fuzzy-System zu erlernen, die sich i.w. in drei Klassen einteilen lassen. Clusterbasierte und hyperquaderorientierte Ansätze erlernen die Regeln (Struktur) und die Fuzzy-Mengen (Parameter) des Fuzzy-Systems gleichzeitig, während strukturorientierte Ansätze eine (Anfangs-)Fuzzy-Partitionierung der Wertebereiche der Variablen benötigen, um eine Regelbasis zu erzeugen [17].

Clusterbasierte Regellernansätze gehen, wie der Name schon nahelegt, von einer Fuzzy-Clusteranalyse der Daten aus [3, 11]. Das bedeutet, daß es sich um unüberwachte Lernverfahren handelt, d.h. es gibt keine vorgegebene Zielgröße. Hyperquaderorientierte Ansätze sind dagegen überwachte Lernverfahren, bei denen man versucht, die Trainingsdaten durch (ggf. überlappende) Hyperquader abzudecken, um so die Abhängigkeit der Zielgröße von anderen Variablen durch *Fuzzy-Graphen* zu beschreiben [2]. In beiden Fällen werden die Fuzzy-Mengen aus der Projektion der Cluster bzw. der Hyperquader auf die Wertebereiche der einzelnen Variablen gewonnen.

Das Hauptproblem dieser Ansätze ist, daß jede erzeugte Fuzzy-Regel eigene spezifische Fuzzy-Mengen verwendet und deshalb (siehe oben) die Regelbasis u.U. nur schwer zu interpretieren ist. Clusterbasierte Verfahren kranken außerdem daran, daß durch die Projektion auf die einzelnen Variablen Information über die u.U. nicht achsenparallele Form der Cluster verloren geht, und daß man mit ihnen eine passende Anzahl von Regeln gewöhnlich nur bestimmen kann, indem man sie mehrfach mit unterschiedlich festgelegter Regelbasisgröße ausführt und dann die Ergebnisse bewertet und vergleicht.

Strukturorientierte Ansätze vermeiden diese Nachteile, weil sie nicht nach (hyperellipsoid- oder hyperquaderförmigen) Clustern im Datenraum suchen. Durch die vorgegebene (Anfangs-)Fuzzy-Partitionierung der Wertebereiche wird über den Datenraum ein mehrdimensionales Fuzzy-Gitter gelegt. Aus diesem Gitter wird eine Regelbasis bestimmt, indem die besetzten Gitterzellen ausgewählt und durch Fuzzy-Regeln beschrieben werden. Nachdem so die Regelbasis festgelegt wurde, werden üblicherweise die Fuzzy-Mengen trainiert, um die Leistung des Fuzzy-Systems zu verbessern.

Dieses Lernverfahren für Fuzzy-Regeln wurde ursprünglich in [18] vorgeschlagen. Eine erweiterte Version wird in dem System NEFCLASS (NEuro-Fuzzy-CLASSification) [16] verwendet. NEFCLASS benutzt ein Gütemaß zur Bewertung der gefundenen Fuzzy-Regeln. Auf diese Weise kann die Größe der Regelbasis automatisch bestimmt werden, indem Regeln nach absteigender Güte hinzugefügt werden, bis alle Trainingsbeispiele abgedeckt sind. Mit Hilfe des Gütemaßes wird außerdem das beste Konsequenz für jede Regel bestimmt. Weiter kann die Zahl der Fuzzy-Regeln begrenzt werden, indem man nur die besten Regeln in die Regelbasis aufnimmt. Auch ist es möglich, die Zahl der Regeln und die Zahl der je Regel benutzten Variablen durch „Pruning“ zu verringern.

Um die Fuzzy-Mengen zu optimieren, verwendet NEFCLASS ein einfaches, backpropagation-artiges Verfahren, das durch Lernverfahren für Neuronale Netze inspiriert ist. Daher auch die Bezeichnung „neuro-fuzzy“ für diesen und verwandte Ansätze. Der Algorithmus führt jedoch nicht, wie das normale Backpropagation-Verfahren, einen Gradientenabstieg durch, da der Erfüllungsgrad einer Regel über das Minimum bestimmt

wird und außerdem die Fuzzy-Mengen oft durch nicht überall differenzierbare Funktionen beschrieben werden. Stattdessen wird eine einfache Heuristik benutzt, durch die die Fuzzy-Mengen verschoben und in ihren Formen verändert werden.

Ein wesentliches Ziel von NEFCLASS ist es, die Verständlichkeit der gelernten Fuzzy-Systems sicherzustellen. Daher wird verhindert, daß die Fuzzy-Mengen durch das Lernverfahren beliebig verändert werden. Verschiedene Arten von Beschränkungen können vorgegeben werden, damit die Fuzzy-Mengen auch nach dem Lernprozeß noch zu den ihnen zugeordneten sprachlichen Begriffen passen. Z.B. sollten benachbarte Fuzzy-Mengen nicht ihre Positionen vertauschen, sie sollten sich angemessen überlappen usw.

Die neueste Java-Implementierung von NEFCLASS hat folgende Eigenschaften:

- struktur-orientiertes Lernen von Fuzzy-Regeln,
- automatische Bestimmung der Regelanzahl,
- Behandlung fehlender Werte (ohne Einsetzen von Schätzwerten),
- Verarbeitung sowohl von numerischen als auch von symbolischen Attributen,
- auswählbare Beschränkungen der Veränderung der Fuzzy-Mengen, und
- automatische Pruning-Strategien.

Das Programm heißt NEFCLASS-J und kann kostenlos von unserer Internetseite geladen werden (`http://fuzzy.cs.uni-magdeburg.de`).

4 Ausblick

Im Bereich der Wissensentdeckung in Datenbanken und des Data Mining gibt es eine Tendenz, sich im ersten Schritt auf rein datengetriebene Ansätze zu beschränken. Stärker modellbasierte Ansätze werden meist nur in Verfeinerungsphasen verwendet (die in der Industrie jedoch oft nicht notwendig sind, da der erste erfolgreiche Versuch gewinnt — und „the winner takes all“). Um jedoch zu wirklich nutzbringenden Ergebnissen zu gelangen, muß man Hintergrundwissen und i.a. auch nicht-numerische Information berücksichtigen und sich auf verständliche Modelle konzentrieren.

Dabei führt die Komplexität der Lernaufgabe offenbar zu einem Problem: Gewöhnlich muß man wählen zwischen (oft quantitativen) Methoden, die eine hohe Performanz erzielen, und (oft qualitativen) Modellen, die einem Anwender Einsichten vermitteln. Dies ist ein (weiteres) gutes Beispiel für Zadeh's Prinzip der Inkompatibilität von Genauigkeit und Verständlichkeit. Natürlich sind Genauigkeit und hohe Performanz wichtige Ziele. Aber in den erfolgreichsten Fuzzy-Anwendungen in der Industrie, wie z.B. intelligente Regelung und Mustererkennung, war die Einführung von Fuzzy-Mengen motiviert durch einen Bedarf an benutzerfreundlicheren Methoden, die einem Anwender helfen, sein Wissen zu formulieren und die verfügbaren Informationen in möglichst einfacher Weise abzurufen, zu strukturieren und zu verarbeiten. Um diese Benutzerfreundlichkeit zu erreichen, werden oft gewisse Einschränkungen der Performanz und der Lösungsqualität in Kauf genommen. Im Data Mining kann man eine ähnliche Tendenz feststellen.

Wünschenswert wäre eine Theorie der Nützlichkeit, die die Einfachheit und Verständlichkeit eines Systems berücksichtigt. Unglücklicherweise ist eine solche Theorie nicht leicht zu formulieren, da es schwierig ist, den Grad der Einfachheit und Verständlichkeit eines Systems zu messen und noch schwieriger den dadurch erreichten Gewinn einzuschätzen. Dennoch ist der Wunsch nach einer solchen Theorie eine bleibende Herausforderung für die Forschung im Bereich der Fuzzy-Systeme.

Literatur

[1] H. Bandemer and W. Näther. *Fuzzy Data Analysis*. Kluwer, Dordrecht, Netherlands 1992

[2] M. Berthold and K.P. Huber. Constructing Fuzzy Graphs from Examples. *Int. J. Intelligent Data Analysis* 3, 1999. (electronic journal: `http://www.elsevier.com/locate/ida`)

[3] J.C. Bezdek, J. Keller, R. Krishnapuram, and N. Pal. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Kluwer, Norwell, MA, USA 1998

[4] C. Borgelt, J. Gebhardt, and R. Kruse. Chapter F1.2: Inference Methods. In: E. Ruspini, P. Bonissone, and W. Pedrycz, eds. *Handbook of Fuzzy Computation*. Institute of Physics Publishing Ltd., Bristol, United Kingdom 1998

[5] P. Chapman, J. Clinton, T. Khabaza, T. Reinartz, and R. Wirth. *The CRISP-DM Process Model*. 1999 (available from `http://www.ncr.dk/CRISP/`)

[6] D. Dubois, H. Prade, and R.R. Yager. Information Engineering and Fuzzy Logic. *Proc. 5th IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE'96, New Orleans, LA, USA)*, 1525–1531. IEEE Press, Piscataway, NJ, USA 1996

[7] D. Dubois, H. Prade, and R.R. Yager. Merging Fuzzy Information. In: J.C. Bezdek, D. Dubois, H. Prade, eds. *Approximate Reasoning and Fuzzy Information Systems, (Series: Handbook of Fuzzy Sets)*, 335–402. Kluwer, Dordrecht, Netherlands 1999

[8] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds. *Advances in Knowledge Discovery and Data Mining*. MIT Press, Cambridge, MA, USA 1996

[9] J. Gebhardt and R. Kruse. Parallel Combination of Information Sources. In: D. Gabbay and P. Smets, eds. *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, Vol. 3:329–375. Kluwer, Dordrecht, Netherlands 1998

[10] P. Gentsch. *Data Mining Tools: Vergleich marktgängiger Tools*. WHU Koblenz, Germany 1999

[11] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy Cluster Analysis*. J. Wiley & Sons, Chichester, England 1999

[12] R. Kruse, J. Gebhardt, and F. Klawonn. *Foundations of Fuzzy Systems*. J. Wiley & Sons, Chichester, England 1994

[13] R. Kruse and K.D. Meyer. *Statistics with Vague Data*. Reidel, Dordrecht, Netherlands 1987

[14] R. Kruse, C. Borgelt, and D. Nauck. Fuzzy Data Analysis: Challenges and Perspectives. *Proc. 8th IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE'99, Seoul, Korea)*. IEEE Press, Piscataway, NJ, USA 1999 (to appear)

[15] G. Nakhaeizadeh. Wissensentdeckung in Datenbanken und Data Mining: Ein Überblick. In: G. Nakhaeizadeh, ed. *Data Mining: Theoretische Aspekte und Anwendungen*, 1–33. Physica-Verlag, Heidelberg, Germany 1998

[16] D. Nauck, F. Klawonn, and R. Kruse. *Foundations of Neuro-Fuzzy Systems*. J. Wiley & Sons, Chichester, England 1997

[17] D. Nauck and R. Kruse. Chapter D.2: Neuro-fuzzy Systems. In: E. Ruspini, P. Bonissone, and W. Pedrycz, eds. *Handbook of Fuzzy Computation*. Institute of Physics Publishing Ltd., Bristol, United Kingdom 1998

[18] L.X. Wang and J.M. Mendel. Generating fuzzy rules by learning from examples. *IEEE Trans. Systems, Man, Cybernetics* 22(6):1414–1427. IEEE Press, Piscataway, NJ, USA 1992