MIDAS – ein Multistrategiesystem zum explorativen Data Mining

Stefan Sklorz, Andreas Becks, Matthias Jarke
RWTH Aachen, Lehrstuhl Informatik V
Ahornstr. 55, D-52056 Aachen
{sklorz, becks, jarke}@informatik.rwth-aachen.de

Zusammenfassung

Viele Anwendungen der Wissensentdeckung benötigen einen explorativen Data Mining-Ansatz, der menschliche Intuition und Sachkompetenz mit effizienten Techniken zur Strukturerkennung und Beschreibung kombiniert. Wir haben mit MIDAS ein System entwickelt, das zu diesem Zweck zwei Integrationsaspekte realisiert: (a) Die Verbindung zwischen der Erkennung von Strukturen durch ein neuronales Netz und ihrer Beschreibung durch ein maschinelles Lernverfahren, sowie (b) die Einbindung des Benutzers in die maschinelle Strukturerkennung und Beschreibung über Visualisierungstechniken. In diesem Beitrag konzentrieren wir uns auf Strukturerkennung und stellen hierzu zwei Basis-Techniken der MIDAS-Datenvisualisierung vor. Erfahrungen mit drei Anwendungsfeldern illustrieren ihren Nutzen und ihre systemtechnische Umsetzung: Eine Analyse medizinischer Labordaten, eine Standortanalyse für den Einzelhandel sowie eine Untersuchung von Strukturen in textuellen Dokumentensammlungen.

1 Einleitung

Auf der Grundlage des Data Warehousing kommt auch dem Data Mining eine wachsende Bedeutung zu. Die integrative Behandlung eines umfassenden Entdeckungsprozesses, in dem sich das Instrumentarium verschiedener Disziplinen nutzbringend einbettet, hat inzwischen eine Vielzahl an Verfahrensweisen für das Data Mining hervorgebracht. Das Spektrum vorhandener Techniken reicht von einer verifikationsgetriebenen Datenanalyse, die etwa durch OLAP-Ansätze und Visualisierungstechniken vertreten wird, über den Einsatz von statistischen Methoden bis hin zu einer betont datengetriebenen Form der Analyse, in der die Leistung des menschlichen Analysten primär in die Datenauswahl und die Hypothesenverifikation eingeht, während die Stärken der eingesetzten Techniken, so etwa vieler maschineller Lernverfahren, vorrangig in der Hypothesenbildung liegen.

Aus anfänglichen „Insellösungen“, die zumeist für Einzelanwendungen entworfen wurden und oft nur wenige Data Mining-Techniken implementierten, haben sich Multistrategiesysteme entwickelt, wie etwa KEPLER [16], Clementine [7] und DataEngine [8]. Solche Systeme verfügen heute über ein ganzes Arsenal verschiedener Techniken und setzen direkt auf den Datenquellen auf. Einerseits gelingt auf diesem Wege eine stärkere Unterstützung des gesamten Entdeckungsprozesses, der von einer Bereichsanalyse über die Datenbeschaffung bis hin zur Interpretation der erzielten Resultate reicht. Andererseits treten bei der Systementwicklung auch neue Probleme in den Vordergrund, die weit in den Bereich des Software-Engineering reichen. Erweiterbare Architekturen [34], die einen Austausch integrierter Techniken anforderungsspezifisch ermöglichen, sowie die Forderung nach einer leichten Anbindung an bestehende Warehouse-Architekturen können dazu führen, daß die für eine systemtechnische Realisierung betriebenen Anstrengungen die Entwicklung der Techniken dominieren.

Unser Ansatz geht dagegen von der These aus, daß auch im geschickten Zusammenwirken von wenigen, unterschiedlichen Techniken noch ungenutzte Potentiale liegen. So eignen sich Neuronale Netze zur Erkennung von Strukturen in numerischen Daten, bieten jedoch *per se* keine Mechanismen zur Erklärung der erkannten Zusammenhänge. Auf Fuzzy Logik basierende Inferenzsysteme liefern dagegen auch auf numerischen Daten vernünftige Erklärungen, ohne jedoch einen Einfluß auf die Bildung der den Folgerungen zugrundeliegenden Regeln und Variablen zu haben. Maschinelle

Lernverfahren können hingegen Regeln zur Strukturbeschreibung lernen, sofern eine geeignete Repräsentation der numerischen Daten gegeben ist. Letztlich bieten vor allem Visualisierungstechniken als Alternative zu Zahlen und Regeln einen leichten Zugang zu den Daten, der bereits ein Gefühl für deren innere Strukturen und Auffälligkeiten vermittelt und dabei erste Fragen aufwirft, deren Beantwortung im Detail die weiteren Untersuchungen leiten kann.

1.1 Data Mining mit MIDAS

Das Data Mining-System MIDAS (Multi-Purpose Interactive Data Analysis System) kombiniert Techniken aus den angesprochenen Bereichen, um auf diesem Wege ihre jeweiligen Beschränkungen als Aufsatz für ein Datenbanksystem zu überwinden [10].

MIDAS unterstützt die Exploration von multidimensionalen numerischen Merkmalsvektoren. Hierbei liegt seine Stärke in einem Verfahren zur unüberwachten Generierung von Fuzzy-Termen [28]. Die Terme werden datengetrieben unter Nutzung der Generalisierungseigenschaft einer neuronalen Merkmalskarte [17] berechnet, nachdem Gruppen ähnlicher Muster in den Daten entdeckt werden konnten. Der Strukturentdeckung dienen algorithmische Segmentierungsverfahren zur Erzeugung hierarchischer Gruppierungen [27] und Visualisierungstechniken, deren Grundlage der mit den Merkmalskarten erzielte Daten- und Strukturgewinn ist. Gemeinsam mit den entdeckten Gruppen bilden die Terme dann eine geeignete Repräsentation der Daten und ermöglichen somit einen Einsatz von maschinellen Lernverfahren, die *Fuzzy-IF-Then*-Regeln zur Beschreibung der gefundenen Strukturen lernen. Bislang wird eine auf MIDAS angepaßte Variante des Lernverfahrens *Fuzzy-ID3* [32] eingesetzt. Mit Hilfe farbiger „Overlays“ können dann gebildete Terme und gelernte Regeln über die in MIDAS realisierten Datenvisualisierungen dargestellt werden. Dies spannt einen Bogen zwischen daten- und verifikationsgetriebener Analyse und unterstreicht zugleich den explorativen Anspruch unseres Ansatzes.

Darauf folgend dienen die Datenvisualisierungen dem Analysten als Plattform für weitere Datenzugriffe, Datenmanipulationen und Experimente. Der Einsatz visueller Darstellungen in MIDAS geht somit über den von rein „statischen“ Abbildungen hinaus. Die hierzu verwendeten Visualisierungstechniken suggerieren dem Betrachter eine Vorstellung von der Struktur in den Daten, lassen Eigenarten und Auffälligkeiten hervortreten und regen ihn zur Bildung von weiteren Hypothesen an, die er unmittelbar anhand der grafischen Darstellungen mit Hilfe geeigneter Operationen überprüfen und so mit den datengetriebenen Resultaten verbinden kann.

1.2 Zielsetzung und Aufbau des Beitrages

Wir illustrieren anhand des MIDAS-Systems das Potential einer Verknüpfung zwischen datengetriebener und verifikationsgetriebener Analyse. Dabei konzentrieren wir uns auf den Aspekt der Benutzereinbindung und stellen hierzu zwei Basis-Techniken der MIDAS-Datenvisualisierung vor.

Abschnitt 2 greift kurz das Modell der Merkmalskarten auf und führt in unsere Notation ein. In Abschnitt 3 betrachten wir ein einführendes Anwendungsbeispiel, das den mit den Merkmalskarten verbundenen Daten- und Strukturgewinn plausibel macht und zur Illustration der im Abschnitt 4 vorgestellten Visualisierungstechniken dient. Danach (Abschnitt 5) diskutieren wir Praxiserfahrungen mit MIDAS, bevor wir in Abschnitt 6 verwandte Arbeiten aus Bereich der Datenvisualisierung betrachten.

2 Merkmalskarten: Modell und Notation

Kohonen’s Merkmalskarten ([17], [18]) sind zweischichtige neuronale feed-forward Netze mit einer Eingabe- und einer Ausgabeschicht von Verarbeitungseinheiten. Jede Eingabeeinheit steht mit jeder Ausgabeeinheit in Verbindung. Andere Verbindungen existieren nicht. Die Ausgabeeinheiten werden

jedoch über eine Nachbarschaftsrelation miteinander in Beziehung gesetzt, die über eine Zuordnung zwischen Ausgabeeinheiten und Punktkoordinaten in einem Raum regelmäßig angeordneter Punkte definiert ist. Üblicherweise handelt es sich dabei um Kreuzungspunkte regelmäßiger Gitterstrukturen. Wir nutzen ein zweidimensionales Gitter zur Beschreibung einer quadratischen Ausgabeschicht mit $m' \times m'$ Ausgabeeinheiten, die wir durch eine zeilenweise fortlaufende Nummerierung bezeichnen. Damit besitzt jede Ausgabeeinheit $i \in \{1, 2, ..., (m')^2\}$ eine eindeutige Gitterposition (i_1, i_2) mit einer geeigneten Umrechnung $i = \kappa_{m'}(i_1, i_2)$ der Indizes.

Jede Eingabeeinheit dient zur Weiterleitung eines Signales $x \in \mathbb{R}$, das sie allen Ausgabeeinheiten parallel und zeitgleich mit den $(m-1)$ Signalen der übrigen Eingabeeinheiten zuführt. Entsprechend empfängt jede der $(m')^2$ Ausgabeeinheiten zeitgleich dasselbe Muster von m Signalen über ihre m Verbindungen zur Eingabeschicht. Jede Verbindung verfügt über einen lokalen Speicher, in dem ein Wert $z \in \mathbb{R}$ abgelegt ist. Wir fassen die bei i gespeicherten Werte $z_{i_1}, z_{i_2}, ..., z_{i_m}$ zu einem *Zustandsvektor* $\vec{z}_i = (z_{i_1}, ..., z_{i_m})^T \in \mathbb{R}^m$ und die eintreffenden Signale $x_1, x_2, ..., x_m$ zu einem *Merkmalsvektor* $\vec{x} = (x_1, ..., x_m)^T \in \mathbb{R}^m$ zusammen. Die Verarbeitung von $\vec{x}$ durch i erfolgt dann durch die Berechnung einer Ausgabe $y = \eta(\vec{x}, \vec{z}_i) \in \mathbb{R}$, wobei η ein Distanz- oder alternativ ein Ähnlichkeitsmaß realisiert. Mit η wird somit die Reaktion einer Ausgabeeinheit i auf $\vec{x}$ beschrieben (*Aktivierungsfunktion*), oder anders formuliert, der Grad ihrer durch $\vec{x}$ ausgelösten „Erregung“ mit maximaler Erregung bei $\vec{x} = \vec{z}_i$ bestimmt. Die insgesamt stärkste Erregung definiert stets eine Ausgabeeinheit c (*Erregungszentrum*), deren Gitterposition die Basis für einen Lernschritt bildet.

In diesem Zusammenhang entspricht „Lernen“ einer Adaption der gespeicherten Information an eine Menge X von Merkmalsvektoren, die in zufälliger Folge und in diskreten Zeitschritten t der Eingabeschicht separat zugeführt werden. Die Zustandsvektoren als Träger der Information werden dabei mit dem Ziel verändert, die wechselseitigen Ähnlichkeiten der Merkmalsvektoren in der über η berechneten Erregung der Ausgabeschicht widerzuspiegeln. Über das Erregungszentrum $c(t)$ eines im Zeitschritt t angelegten Merkmalsvektors $\vec{x}(t)$ wird hierzu eine zeitabhängige Nachbarschaft $\varphi_{ci}(t)$ definiert, die eine Menge von Ausgabeeinheiten in der Gitterumgebung von c bestimmt. Innerhalb dieser Menge werden alle Zustandsvektoren um einen für jedes i separat bestimmten Vektor $\Delta\vec{z}_i(t)$ in Richtung auf $\vec{x}(t)$ verschoben. Üblicherweise wird $\varphi_{ci}(t)$ mit fortschreitendem t verkleinert, um die Zustandsvektoren allmählich in einen stabilen Zustand zu führen. Zu diesem Zweck fällt auch das Ausmaß der spezifischen Korrekturen $\Delta\vec{z}_i(t)$ mit jedem Lernschritt geringer aus.

In den spezifischen Korrekturen und der Nachbarschaftdefinition bietet Kohonen's Modell viel Raum für individuelle Realisierungen. Theoretische Arbeiten (e.g. [26], [31]) belegen jedoch Eigenschaften, die unter Einhaltung bestimmter Randbedingungen zu erreichen sind. Hierzu zählt die Möglichkeit, die Merkmalsvektoren eines mehrdimensionalen Merkmalsraumes unter bestmöglicher Beibehaltung ihrer Lage- und Anordnungsbeziehungen auf ein Gitter niedrigerer Dimension abzubilden. Dabei bleibt die „Topologie“ des Merkmalsraumes im intuitiven Sinne soweit wie möglich erhalten, d.h. jeder Merkmalsvektor wird durch mindestens einen (nahezu) identischen Zustandsvektor vertreten, und die Zustandsvektoren benachbarter Ausgabeeinheiten beschreiben nach dem Adaptionsprozeß auch benachbarte Punkte im Merkmalsraum. Neben einem Datengewinn verschaffen die Merkmalskarten also auch einen Strukturgewinn.

3 Ein einführendes Beispiel

MIDAS unterstützt die Analyse von Merkmalsvektoren zur Beschreibung „abstrakter“ Objekte, bei denen es sich um beliebige Dinge aus unserem alltäglichen Erfahrungsbereich handeln kann, etwa um Dokumente, Patienten und Filialen. Voraussetzung dabei ist, daß eine Menge X von n Objekten mit jeweils m Merkmalen vorliegt, deren Ausprägungen $\vec{x}_j$ in $\mathbb{R}$ liegen. Ein Merkmalsvektor $\vec{x}_k \in$

X beschreibt dann ein Objekt k, wenn für $\vec{x} \in X$ die gleiche Reihenfolge derselben Merkmale feststeht. Die Ausprägungen ausgewählter Merkmale betrachteter Objekte sind die zu analysierenden Daten, und die Gesamtheit ihrer Ausprägungen spannt den von MIDAS untersuchten und für eine Fragestellung relevanten Merkmalsraum auf.

Abbildung 1a zeigt als Beispiel den Phosphatase- und Eisengehalt im Blutserum von zwanzig Patienten [9]. Mit den Blutproben liegen für jeden Patienten k zwei Merkmale vor, die wir zu einem Merkmalsvektor $\vec{x}_k$ zusammenfassen. Jedem $\vec{x}_k$ entspricht ein Punkt in einem (x,y)-Koordinatensystem, an dessen Position eines von vier möglichen „Diagnosesymbolen“ (Normal, Hepatitis, Leberzirrhose, Verschlußikterus) eingetragen ist. Auf diese Weise wird eine Struktur sichtbar: Die $\vec{x}_k$ bilden vier Punkthaufen (*Cluster*) und jedem Punkthaufen ist ein anderer medizinischer Befund zugeordnet.

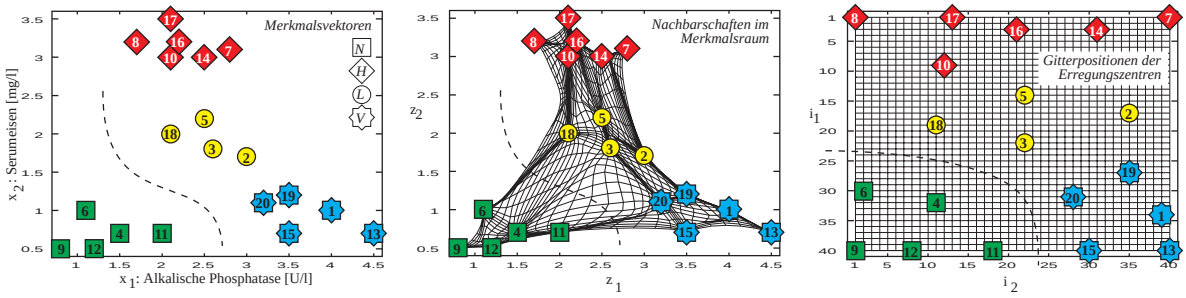

Abbildung 1: *Von links nach rechts*: a) Merkmalsraum mit Merkmalsvektoren und Befunden. b) „Einbettung“ von 1600 adaptierten Zustandsvektoren in den Merkmalsraum. c) „Entfaltung“ auf die Gitterpositionen.

Wir nutzen die $\vec{x}_k$ als Eingangssignale für eine Merkmalskarte und legen sie im Verlauf von 2000 Lernschritten an ein quadratisches Gitter mit 40×40 Ausgabeeinheiten an. Abbildung 1b illustriert das Resultat der erreichten Adaption: Jeder Zustandsvektor beschreibt einen Punkt im Merkmalsraum, wobei je zwei durch $\vec{z}_i \neq \vec{z}_j$ beschriebene Punkte durch eine Linie miteinander verbunden sind, wenn die Gitterpositionen von i und j in genau einer Koordinate übereinstimmen.

Dies entspricht einer „Einbettung“ des Gitters in den Merkmalsraum, deren Bild offensichtlich eine gittersymmetrische Anordnung der Verbindungslinien zeigt. Folglich ist die gefundene Zuordnung zwischen Gitterpositionen und Punkten keineswegs zufällig sondern nachbarschaftserhaltend: Benachbarte Ausgabeeinheiten speichern über ihre Zustandsvektoren auch benachbarte Punkte im Merkmalsraum. Die $\vec{z}_i$ sind also nach Ähnlichkeit geordnet. Dabei ist jeder Merkmalsvektor durch mindestens einen ihm äußerst ähnlichen Zustandsvektor wiedergegeben. Nummern kennzeichnen die jeweiligen Erregungszentren. Relativ zueinander betrachtet ergeben daher die Gitterpositionen der Erregungszentren eine Art „Spiegelbild“ der Merkmalsraum-Positionen der Merkmalsvektoren.

Dies illustriert Abbildung 1c. Hierzu legen wir die Punkte aus Abbildung 1b unter Beibehaltung ihrer Verbindungslinien und der markierten Erregungszentren auf die Koordinaten ihrer Ausgabeeinheiten. Ein Vergleich mit Abbildung 1a zeigt, daß sich die Nachbarschaften der $\vec{x}_k$ in den Gitterpositionen der $c_{\vec{x}_k}$ wiederfinden lassen. Einen eindeutigen Zusammenhang zwischen den Abständen der Gitterpositionen und den Ähnlichkeiten der Merkmalsvektoren gibt es jedoch nicht. Aufgrund einer weitgehend gleichmäßigen Verteilung der $c_{\vec{x}_k}$ ist eine Ordnung nur anhand ihrer nach Diagnosen aufgeschlüsselten Markierungen erkennbar. Dies läßt sich darauf zurückführen, daß die Merkmalskarten auch die Häufigkeitsverteilung der $\vec{x}_k$ berücksichtigen: Je mehr Merkmalsvektoren in einer bestimmten Region im Merkmalsraum liegen, desto mehr Zustandsvektoren werden für diese Region gebildet. Entsprechend liegen auch im Beispiel deutlich mehr Zustandsvektoren in den Umrissen der vier Punkthaufen als zwischen diesen Bereichen, was in Abbildung 1b einer durch mehr Gitterpunkte beschriebenen Wiedergabe dieser Regionen entspricht.

4 Basis-Techniken der MIDAS-Datenvisualisierung

Der mit den Merkmalskarten verbundene Daten- und Strukturgewinn bildet die Basis der MIDAS-Datenvisualisierung. Wir haben darauf aufbauend eine Reihe von aufgabenspezifischen Techniken zur Datenanalyse in MIDAS realisiert, von denen wir zwei anhand der Beispieldaten vorstellen.

4.1 Erregungsmuster einzelner Objekte

Die stärkste Erregung $\eta(\vec{x}, \vec{z}_c)$ definiert für jeden Lernschritt ein Erregungszentrum c, um damit die Adaption von Zustandsvektoren an $\vec{x}$ einzuleiten. Hierzu quantifiziert η für jede Ausgabeeinheit i den Grad der Ähnlichkeit bzw. den Grad der Verschiedenheit zwischen $\vec{z}_i$ und $\vec{x}$ durch eine Zahl $\eta(\vec{x}, \vec{z}_i) \in \mathbb{R}$. Gemeinsam formen diese Zahlen ein durch den angelegten Merkmalsvektor $\vec{x}$ in der Ausgabeschicht hervorgerufenes *Erregungsmuster* $\mathcal{A}^{\vec{x}}$. Für ein quadratisches Gitter mit $(m')^2$ Ausgabeinheiten läßt sich $\mathcal{A}^{\vec{x}}$ nach Gl. 1 in einer $m' \times m'$-Matrix speichern. Mit der Umrechnung $i = \kappa_{m'}(i_1, i_2)$ entspricht ein Matrixeintrag $\mathcal{A}^{\vec{x}}_{i_1 i_2}$ dabei der Erregung $\eta(\vec{x}, \vec{z}_i)$, also der Erregung, die $\vec{x}$ an der Gitterposition (i_1, i_2) bei der Ausgabeeinheit i auslöst.

$$\mathcal{A}^{\vec{x}}_{i_1 i_2} := \eta(\vec{x}, \vec{z}_i) \quad \text{für alle} \quad 1 \leq i_1, i_2 \leq m';\ i = \kappa_{m'}(i_1, i_2);\ \vec{x} \in X \tag{1}$$

Im Lernprozeß unterliegen die Erregungsmuster starken Veränderungen. Anfänglich noch zufällige Matrixeinträge weichen allmählich musterspezifischen Strukturen, die durch Felder ähnlicher Werte in verschiedener Form und Ausdehnung gekennzeichnet sind. Dabei bilden sich Eigentümlichkeiten und Besonderheiten heraus, die als Folge der zunehmenden Ordnung der Zustandsvektoren über einen paarweisen Vergleich wechselseitige Ähnlichkeiten von Merkmalsvektoren aufdecken. So verursachen ähnliche Merkmalsvektoren stets eine Reihe von gemeinsamen „Erkennungsmarken“, die mit Hilfe einer geeigneten Mustervisualisierung eine sinnvolle Gruppierung von *a priori* nicht klassifizierten Objekten erlauben. Hierzu bieten Graustufenbilder eine Möglichkeit, die den Gitterpositionen eine weitere Dimension geben. Abbildung 2 illustriert dies am Beispiel.

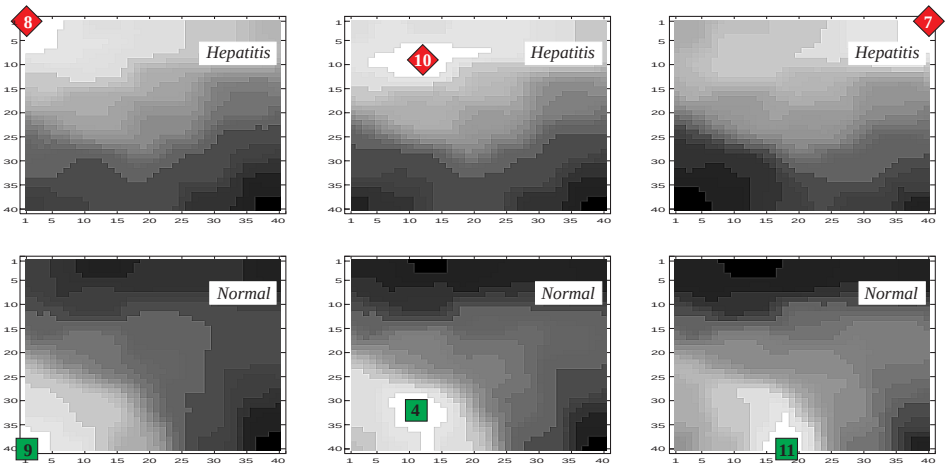

Abbildung 2: Durch Graustufenbilder visualisierte Erregungsmuster. Von links: $\mathcal{A}^{\vec{x}_8}$, $\mathcal{A}^{\vec{x}_{10}}$, $\mathcal{A}^{\vec{x}_7}$, $\mathcal{A}^{\vec{x}_9}$, $\mathcal{A}^{\vec{x}_4}$, $\mathcal{A}^{\vec{x}_{11}}$. Je zwei Bilder einer Reihe wirken optisch ähnlicher als beliebige Paare beider Reihen.

In Abbildung 2 sind die Matrixeinträge der Beispielmuster nach dem Lernprozeß ihren Positionen entsprechend durch graustufige Bildpunkte auf 40×40-Rastern dargestellt (in jedem Bild linear skaliert mit Maximum auf schwarz und Minimum auf weiß bei 64 Graustufen). Dabei realisiert η ein Di-

stanzmaß. Besonders helle Felder um die Erregungszentren markieren daher Ausgabeeinheiten, deren Zustandsvektoren fast identisch mit den jeweils angelegten Merkmalsvektoren sind. Auch wenn diese Darstellung noch recht einfach ist, sind die an Hepatitis erkrankten Patienten einerseits und die Patienten mit normalem Befund andererseits durch Graustufenverläufe in den Mustern gekennzeichnet, die in ihrer Gesamtheit einen visuellen Eindruck von „Ähnlichkeit“ in der jeweiligen Befundungsgruppe vermitteln. Die Visualisierung ermöglicht also Betrachtern eine intuitive Objektgruppierung, die den wechselseitigen Ähnlichkeiten der objekt-beschreibenden Merkmalsvektoren folgt.

Von auftretenden Dimensionskonflikten bei der Erzeugung der Merkmalskarten abgesehen, ist die Mustervisualisierung unabhängig von der Datendimension, d.h. der Anzahl der Merkmale pro Objekt. Je nach Zählweise bleibt es bei den zwei bzw. den drei benötigten Visualisierungsdimensionen. Somit ist prinzipiell die Untersuchung von hochdimensionalen Merkmalsräumen möglich. Die durch den visuellen Bildeindruck gewonnene „Trennschärfe“ der Objekte entspricht dabei der Qualität der mit den Merkmalskarten im Einzelfall erreichbaren Topologieerhaltung.

4.2 Überlagerte Erregungsmuster

Erregungsmuster einzelner Objekte leisten eine Dimensionsreduktion der Daten. Diese unterstützt einen Analysten in einer Datenreduktion, und zwar der Zusammenfassung von ähnlichen Objekten. Für eine große Anzahl von Objekten bleibt diese Unterstützung jedoch unbefriedigend. Hunderte von Untersuchungsobjekten konfrontieren einen Analysten mit ebenso vielen Mustervisualisierungen — eine Menge, die kaum zu handhaben und noch weniger zu interpretieren ist. Abhilfe schafft die Bildung von *überlagerten* Erregungsmustern $\mathcal{P}^M$ nach Gl. 2.

$$\mathcal{P}^M_{i_1 i_2} := \min_{\vec{x} \in M} \{\mathcal{A}^{\vec{x}}_{i_1 i_2}\} \quad \text{für alle} \quad 1 \leq i_1, i_2 \leq m';\ M \subseteq X;\ |M| > 1 \tag{2}$$

Für eine zwei- oder mehr-elementige Teilmenge M der Menge der Merkmalsvektoren X definiert eine Überlagerung der für alle $\vec{x} \in M$ hervorgerufenen Erregungsmuster $A^{\vec{x}}$ ein Erregungsmuster $\mathcal{P}^M$, das sich aus den stärksten Einzelerregungen aller Ausgabeeinheiten für alle $\vec{x} \in M$ zusammensetzt. Je nachdem, ob die Aktivierungsfunktion η eine Distanz- oder eine Ähnlichkeitsfunktion realisiert, entspricht die stärkste Erregung einer Ausgabeeinheit i dabei dem Minimum oder dem Maximum der Werte $\eta(\vec{x}, \vec{z}_i)$ für alle $\vec{x} \in M$. Entsprechend berechnet Gl. 2 das überlagerte Erregungsmuster $\mathcal{P}^M$ für eine Distanzfunktion η und ein quadratisches Gitter mit $(m')^2$ Ausgabeeinheiten. Ein Matrixeintrag $\mathcal{P}^M_{i_1 i_2}$ entspricht dabei der kleinsten Distanz aller paarweisen Distanzen zwischen dem Zustandsvektor z_i einer Ausgabeeinheit i an Gitterposition (i_1, i_2) und allen $\vec{x} \in M$.

Die Berechnung von überlagerten Erregungsmustern erscheint zunächst wenig nützlich: Aus 2^n Teilmengen einer Menge X von n Merkmalsvektoren ergeben sich $2^n - (n+1)$ verschiedene Muster, deren Betrachtung selbst für ein kleines n ausgeschlossen bleibt. Benötigt wird also eine Strategie zur Auswahl von Mustern. Hierzu illustriert Abbildung 3 eine bestechend einfache aber ebenso wirkungsvolle Heuristik, die anhand der folgenden Betrachtungen plausibel wird.

Für ein beliebiges $A^{\vec{x}_j}$ (z.B. $A^{\vec{x}_8}$ aus Abb. 2) markieren besonders dunkle Felder in seinem Graustufenbild alle $\vec{z}_i$ mit großer Distanz zu $\vec{x}_j$. Dabei entsprechen den dunkelsten Feldern die größten Distanzen. Weiterhin liegt in jedem dunklen Feld mindestens ein Erregungszentrum $c_{\vec{x}_k}$ (vgl. Abb. 2 mit Abb. 1c). Dies folgt aus dem Bildungsprozeß der Merkmalskarten und der Ordnung der Zustandsvektoren. Nach Definition entspricht die Distanz zwischen $\vec{x}_k$ und $\vec{z}_{c_{\vec{x}_k}}$ dabei der kleinsten Distanz aller paarweisen Distanzen zwischen $\vec{x}_k$ und allen $\vec{z}_i$. Aufgrund der durchgeführten Adaption ist zudem plausibel, daß alle $\vec{x} \in X$ durch (nahezu) identische Zustandsvektoren vertreten sind. Demnach liegt obige Distanz nahe bei Null. Dann folgt aber auch für das Erregungszentrum $c_{\vec{x}_k}$ im „dunkelsten“ Feld aller dunklen Felder in $A^{\vec{x}_j}$ (gibt es mehrere, dann in einem davon; für $A^{\vec{x}_8}$ ist dies $c_{\vec{x}_{13}}$), daß die

Distanz zwischen $\vec{x}_j$ und $\vec{x}_k$ gleich der maximalen Distanz aller paarweisen Distanzen zwischen $\vec{x}_j$ und allen $\vec{x}$ aus X ist (also $\vec{x}_{13}$ für $\vec{x}_8$; vgl. mit Abb. 1a).

Dies macht $\mathcal{P}^{M_1}$ mit $M_1 = \{\vec{x}_j, \vec{x}_k\}$ zu einer guten Wahl, denn $\vec{x}_j$ und $\vec{x}_k$ werden aller Voraussicht nach verschiedenen Punkthaufen angehören (so auch $\vec{x}_8$ und $\vec{x}_{13}$), sofern tatsächlich Strukturen dieser Art in X enthalten sind. Nehmen wir an, es gibt genau zwei entsprechende Punkhaufen in X. Dann wird das Graustufenbild von $\mathcal{P}^{M_1}$ den visuellen Eindruck vermitteln, daß es sich in zwei hellere Bereiche aufteilt, die durch einen dunkleren und schmaleren „Trennstreifen“ abgegrenzt sind (vgl. mit Abb. 3a: im Beispiel sind es 4 Punkthaufen, und das zusätzliche dunkle Teilfeld ergibt aus einem, der abseits der beiden betrachteten liegt). Die Trennung verläuft dabei zwischen $c_{\vec{x}_j}$ und $c_{\vec{x}_k}$ und führt über die Gitterpositionen der übrigen Erregungszentren zu einer Aufteilung von X, die den beiden Punkthaufen entspricht. Dies ist aufgrund der Ordnung der $\vec{z}_i$ und ihrer Verteilung im Merkmalsraum plausibel. Die Qualität der Ordnung bestimmt dabei die Stärke der visuellen Trennung im Bild.

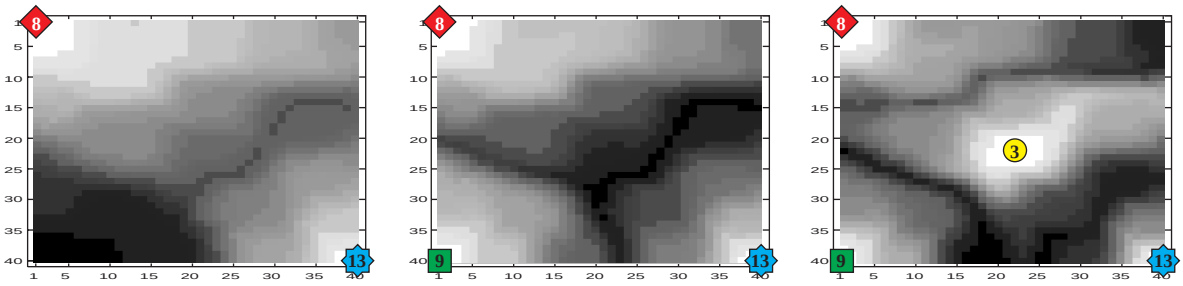

Abbildung 3: $\mathcal{P}^M$-Matrizen des Beispiels. *Von links:* a) $\mathcal{P}^{M_1}$. b) $\mathcal{P}^{M_2}$. c) $\mathcal{P}^{M_3}$. Die schrittweise Überlagerung nach unserer Heuristik bildet die gesuchten Strukturen nach zwei Erweiterungen heraus.

Dieselbe Argumentation läßt sich leicht für eine beliebige Anzahl von Punkthaufen fortsetzen. In diesem Fall erweitern wir die Menge M_1 schrittweise zu Mengen $M_2 \subset M_3 \subset M_4 \subset \ldots$. Dabei entsteht M_{i+1} aus M_i durch die Hinzunahme eines der Merkmalsvektoren, deren Erregungszentren in einem der dunkelsten Felder von $\mathcal{P}^{M_i}$ liegen (für $\mathcal{P}^{M_1}$ in Abb. 3a ist dies $\vec{x}_9$). Falls nun $i+1$ Punkthaufen in X enthalten sind, würde das Graustufenbild von $\mathcal{P}^{M_i}$ der obigen Heuristik folgend bereits nach $i-1$ entsprechenden Erweiterungen in $i+1$ Teilfelder aufgeteilt (vgl. mit Abb. 3c: die Erweiterung von M_1 um $\vec{x}_9$ und $\vec{x}_3$ führt zu $\mathcal{P}^{M_3}$ mit vier Teilfeldern). Die Aufteilung von X erfolgt dann über die Positionen der Erregungszentren in $\mathcal{P}^{M_i}$, wobei jeweils alle Merkmalsvektoren zu einer Gruppe zusammengefaßt werden, deren Erregungszentren in einem abgetrennten Teilfeld liegen. Mögliche Unsicherheiten bei dieser Aufteilung lassen sich zumeist mit wenigen Ergänzungen oder über die vorangegangenen Muster ausräumen (vgl. Abb. 3c mit 3b: nach $\mathcal{P}^{M_3}$ sind $\vec{x}_{19}$, $\vec{x}_{20}$ nicht klar zuzuordnen, wohl aber nach $\mathcal{P}^{M_2}$).

4.3 Überlagerung aller Erregungsmuster

Die in den Vorkapiteln eingeführten $\mathcal{P}^M$-Matrizen stellen eine Verallgemeinerung der $\mathcal{P}$-Matrizen dar, die der ursprünglichen Version von MIDAS zugrundeliegen. Eine $\mathcal{P}$-Matrix entspricht $\mathcal{P}^{M=X}$, also der Überlagerung aller Erregungsmuster in einem Schritt, deren Berechnung sich parallel zur beschriebenen Vorgehensweise anbietet.

Das Graustufenbild von $\mathcal{P}^{M=X}$ macht vor allem eine Trennung zwischen besonders stark ausgeprägten Strukturen im Merkmalsraum deutlich. Klare Strukturen – wie in den Beispieldaten – sind aber eher die Ausnahme und sollten nicht darüber hinwegtäuschen, daß mit der Analyse unbekannter Daten oft eine intensive Auseinandersetzung in der oben beschriebenen Weise verbunden ist.

Abbildung 4a zeigt das Graustufenbild von $\mathcal{P}^{M=X}$ für unser Beispiel, in dem eine starke Trennung zwischen Patienten mit normalem Befund und der Restgruppe zu erkennen ist. In der Restgruppe sind vor allem die Hepatitis-Patienten von den verbleibenden Patienten zu unterscheiden. Obgleich schon schwächer ausgeprägt, wird aber auch eine Aufteilung zwischen den Befunden „Leberzirrhose“ und „Verschlußikterus“ sichtbar. Diese Interpretation führt also zu einer hierarchischen Objektgruppierung, die nur anhand der Graustufenverläufe erfolgt und somit unabhängig von den nach Diagnosen aufgeschlüsselten Markierungen möglich ist.

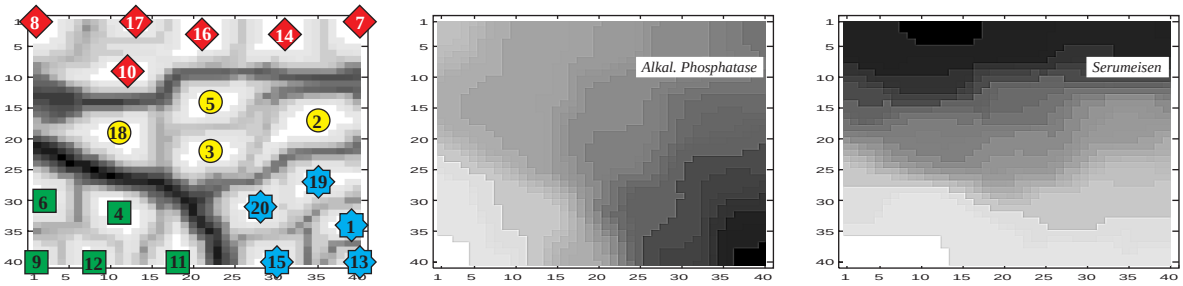

Abbildung 4: *Von links*: a) $\mathcal{P}^M$-Matrix als Überlagerung aller Erregungsmuster. b) Merkmalsverteilung der alkal. Phosphatase. c) Merkmalsverteilung von Serumeisen.

Für diesen Zweck liefern auch die Merkmalsverteilungen in den Zustandsvektoren eine nützliche Information, die wir für jedes Merkmal j nach Gl. 3 in einer $m \times m$-Matrix M^j speichern und durch Graustufenbilder visualisieren können. Die Abbildungen 4b und 4c illustrieren dies am Beispiel.

$$M^j_{i_1 i_2} := z_{i_j} \quad \text{für alle} \quad 1 \leq i_1, i_2 \leq m'; \; i = \kappa_{m'}(i_1, i_2); \; \vec{z}_i = (z_{i_1}, ..., z_{i_m})^T \tag{3}$$

In Verbindung mit Abbildung 4a ergeben 4b und 4c eine Beschreibung der Befundungsgruppen. Verschlußikterus-Patienten unterscheiden sich von allen anderen Patienten durch einen hohen Anteil an alkalischer Phosphatase. Demgegenüber ist eine Hepatitis durch einen hohen Anteil an Serumeisen gekennzeichnet. Niedrige Werte in beiden Merkmale besitzen ausschließlich Patienten mit normalem Befund, während mittlere Werte bei nur bei einer Leberzirrhose auftreten.

5 Anwendungserfahrungen

Das MIDAS-System ist in verschiedenen Anwendungsgebieten des Data Mining mit jeweils anderer Problemstellung eingesetzt worden. Darunter waren mehrere Fallstudien, die wir in enger Kooperation mit namhaften Unternehmen durchführen konnten. Die Resonanz der Anwendungsexperten auf die System-Konzeption und die damit erzielten Resultate ist überwiegend positiv gewesen.

Besonderes Interesse hat dabei stets die Möglichkeit hervorgerufen, die maschinell generierte Fuzzy-Regelmenge zur Beschreibung der Daten mit den vorgestellten Datenvisualisierungen zu verknüpfen. Hierzu wird durch einen beliebigen Satz der erzeugten Regeln für jeden Zustandsvektor $\vec{z}_i$ ein Belief-Wert berechnet, über den dann in Abhängigkeit seiner Größe eine farblich transparente Einfärbung der $\mathcal{P}^M$-Matrizen an der Position (i_1, i_2) vorgenommen wird. Damit läßt sich der Wirkungsgrad einer einzelnen Regel oder einer Regelmenge auf den generalisierten Daten visualisieren und anhand der transparent unterlegten Strukturvisualisierung überprüfen.

Bei den Experten rief dies oft den Ergeiz hervor, über die in MIDAS implementierten interaktiven Analyseoperationen selbst Regeln mit besserer Abdeckung zu formulieren. Hierzu wurden vornehmlich Visualisierungen der Merkmalsverteilungen M^j in Verbindung mit Auswahlboxen verwendet.

Auswahlboxen erlauben die Formulierung von einfachen Und-/Oder-Regeln über Intervalloperationen auf den Merkmalswerten (z.B. `is_greater`, `is_equal`, `is_in[]`, ...), wobei sich die Regeln den Fuzzy-Regeln entsprechend auf die $\mathcal{P}^M$-Matrizen abbilden lassen. Möglich ist aber auch, die vom System generierten Terme und Regeln den eigenen Vorstellungen anzupassen, um somit zu besseren Strukturbeschreibungen zu gelangen. Durch diese Interaktion mit dem System und das damit eingebrachte Expertenwissen konnten die maschinell gelernten Regeln oft zu ersten Hypothesen und folgenden Einsichten ausgebaut werden. Dabei wurde den Ergebnissen eine generell große Akzeptanz entgegengebracht – nicht zuletzt deshalb, weil die Experten am Findungsprozeß aktiv teilnahmen und über die Visualisierungen stets das Gefühl behielten auch größere Regelmengen in ihrer Wirkung noch vollständig zu verstehen.

Wir können in diesem Beitrag nicht auf die Ergebnisse der durchgeführten Studien eingehen. Stattdessen beschreiben wir beispielhaft drei Anwendungsfelder, in denen wir das MIDAS-System eingesetzt haben. Einige Screen-Shots von Teilergebnissen illustrieren den Nutzen der in diesem Beitrag vorgestellten Visualisierungstechniken und ihre systemtechnische Umsetzung.

5.1 Labordatenanalyse

Grundlage medizinischer Befunde bilden mikroskopisch oder durch Meßinstrumente ermittelte Labordaten. Data Mining-Techniken stellen hier in Aussicht, Besonderheiten und Auffälligkeiten zu entdecken, Krankheitsbilder zu klassifizieren und (vielleicht neu und besser) zu beschreiben. Wir haben das MIDAS-System zur Analyse verschiedener Labordatensätze eingesetzt, unter anderem zur Beschreibung von Krebszellen mit Hilfe mikroskopisch ermittelter Zellmerkmale.

Hier illustrieren wir anhand eines Auszuges aus dem *breast cancer Wisconsin Datensatz*[1] [25, 33, 35] den Nutzen der $\mathcal{P}^M$-Matrizen sowie ihr Zusammenspiel mit der maschinell gelernten Regelmenge. Abbildung 5 zeigt das Resultat einer Beispielanalyse.

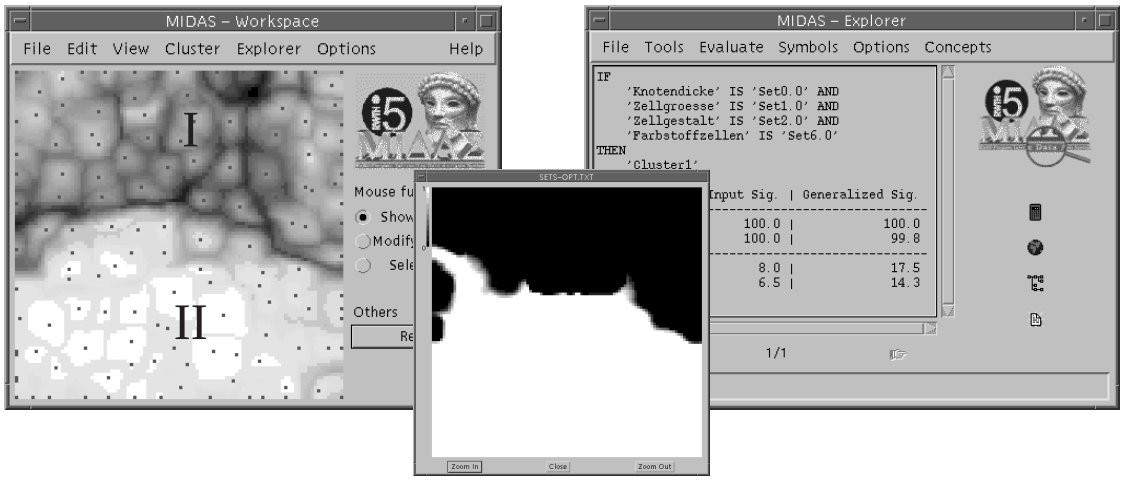

Abbildung 5: *Von links nach rechts*: a) Aufteilung zwischen Zellen ohne Befund (Bereich I) und Tumorzellen (Bereich II). b) Belief-Werte für die Zustandsvektoren. c) Beschreibung der Tumorzellen.

In unserem Beispiel beschreiben die Merkmalsvektoren neun mikroskopisch erhobene Zellmerkmale, wie etwa Zellgröße oder -form, die allesamt auf numerische Wertebereiche abgebildet sind. Anhand der Merkmale ist nach medizinischem Befund eine Trennung zwischen Tumorzellen und

[1] Frei verfügbar über das *UCI Machine Learning Database Repository*: `http://www.ics.uci.edu/`.

normalen Zellen möglich. Aus dem 699 Beschreibungen umfassenden Originaldatensatz haben wir zur Demonstration zufällig 180 Merkmalsvektoren ausgewählt.

Abbildung 5a zeigt das Graustufenbild der Überlagerung aller Erregungsmuster. In diesem Bild ist eine Aufteilung in zwei Bereiche sichtbar. Ein Vergleich mit den medizinischen Befunden ergibt, daß alle Zellen ohne Befund im Bereich I und alle Tumorzellen im Bereich II liegen. Gegenüber Bereich II weist Bereich I noch eine starke Strukturierung in seinen Graustufenverläufen auf, was gleichbedeutend damit ist, daß Zellen ohne Befund gegenüber Tumorzellen in ihren Merkmalen eine größere Variation aufweisen oder umgekehrt, daß Tumorzellen gegenüber Zellen ohne Befund durch eine Reihe gemeinsamer Merkmale beschrieben sind.

Entsprechend findet MIDAS auch nur eine Fuzzy-Regel, um Tumor- von Normalzellen abzugrenzen. Diese Regel ist in Abbildung 5c dargestellt und deckt 100% der Tumorzellen bei einer Fehlklassifikation von 8% korrekt ab. In diese Angaben sind die ermittelten Belief-Werte nicht eingerechnet. Stattdessen zeigt Abbildung 5b für jeden Zustandsvektor im Gitter den Gültigkeitsgrad der gelernten Regel als Grauwert, wobei gilt: Je heller der Bildpunkt, desto stärker „feuert“ die Regel. Alternativ können auch transparente Einfärbungen der $\mathcal{P}^M$-Matrizen vorgenommen werden.

5.2 Standortplanung

Standortplaner versuchen anhand demoskopischer und filialspezifischer Daten mögliche Filialstandorte im Hinblick auf den zu erwartenden Bruttoumsatz zu bewerten. Dabei ist die Güte einer Standortbewertung von großer finanzieller Bedeutung. So laufen Mietverträge in der Regel über 10 bis 15 Jahre, so daß sich eine Fehlbewertung auch über diesen Zeitraum auswirkt. Eine vernünftige Einschätzung muß in jedem Einzelfall die Kombination verschiedener Merkmale berücksichtigen. So können sich in Einzelfällen Merkmale, die für bestimmte Filialen günstig waren, für andere Standorte als ungünstig erweisen. Solche Einzelfälle aufzuspüren und zu beschreiben, ist eine mögliche Aufgabe des Data-Mining.

In Kooperation mit einer großen Einzelhandelskette haben wir das MIDAS-System für diesen Zweck eingesetzt und Datensätze von 816 Filialen mit jeweils 80 Merkmalen untersucht. Um die Vertraulichkeit der Daten zu wahren, beschreiben wir hier lediglich einige allgemeine, aber dennoch interessante Erfahrungen. Hierzu zeigt Abbildung 6 beispielhaft das Resultat einer Untersuchung.

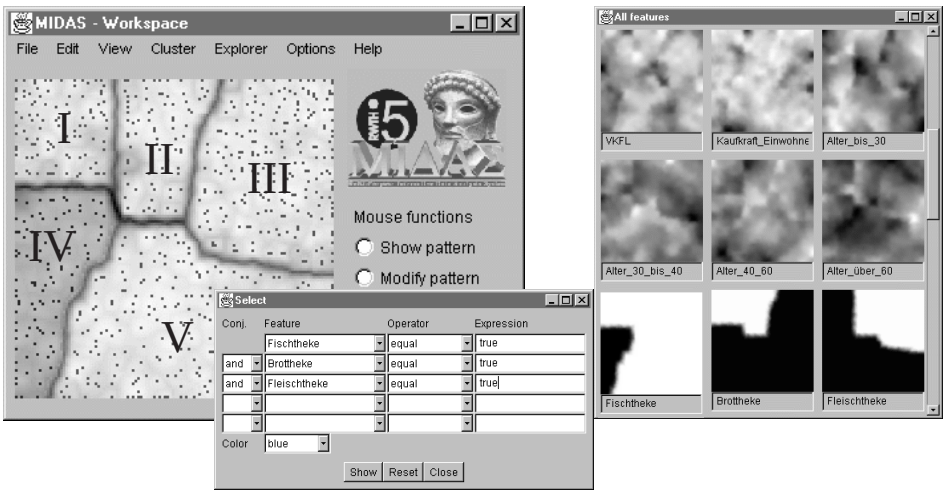

Abbildung 6: *Von links nach rechts*: a) Strukturierung von 816 Filialen. b) Auswahlbox mit Regel zur Beschreibung von Bereich IV. c) Visualisierung einiger Merkmalsverteilungen M^j.

Die täglich in großem Maße anfallenden Daten, die zur Standortanalyse herangezogen werden, umfassen Filialstammdaten zur Charakterisierung der Zweigstelle, Ergebnisdaten, die ihren Erfolg widerspiegeln sowie Angaben über das Sortiment. Demoskopische Daten werden zusätzlich mit diesen Filialinformationen verknüpft. Bei dem Versuch, die Frage nach einem Zusammenhang zwischen Filialerfolg, den angebotenen Serviceleistungen und der Bevölkerungsstruktur mit Hilfe einer Auswahl von 13 relevanten Merkmalen zu beantworten, entdeckt MIDAS fünf Filial-Gruppen (Abb. 6a).

Abbildung 6c gibt deren Merkmalsverteilungen an, wobei helle Graustufen für kleine und dunkle für große Merkmalsausprägungen stehen. Beim Vergleich der Verteilungen mit der erkannten Struktur drängt sich schnell die Vermutung auf, daß die Gruppen lediglich aufgrund des Vorhandenseins oder Nicht-Vorhandenseins von *Brot-*, *Fleisch-* und *Fischtheke* entstanden sind.

Mit Hilfe der Auswahlbox läßt sich leicht eine Regel formulieren, die Filialen mit vielen Serviceleistungen charakterisiert (Abb. 6b). Die Einfärbung des Bereiches IV im Graustufenbild bestätigt die Vermutung. Tiefere Zusammenhänge in den Filialstrukturen können im Beispieldatensatz nicht gefunden werden. Die enge Verknüpfung von maschineller Datenanalyse und Visualisierung in MIDAS hilft somit schnell, uninteressante Muster im Datenmaterial als solche auch zu erkennen.

5.3 Dokumentenlandkarten

Neben der Analyse strukturierter Daten wird in jüngster Zeit auch die inhaltliche Untersuchung von Textsammlungen immer stärker diskutiert. Selbstorganisierende Merkmalskarten werden hier zum Organisieren von Newsgroup-Artikeln [21, 20], als Kategorienkarten für bibliographische Informationssysteme [24] oder zur Klassifizierung von WWW-Homepages [5] eingesetzt. Ein weiteres wichtiges Beispiel für das Anwendungspotential dieser Technologie ist ihr Einsatz als Dokumentenlandkarte für das Wissensmanagement in spezialisierten Textsammlungen – ein Thema, an dem wir gegenwärtig verstärkt arbeiten. Abbildung 7 illustriert ein Beispiel hierfür.

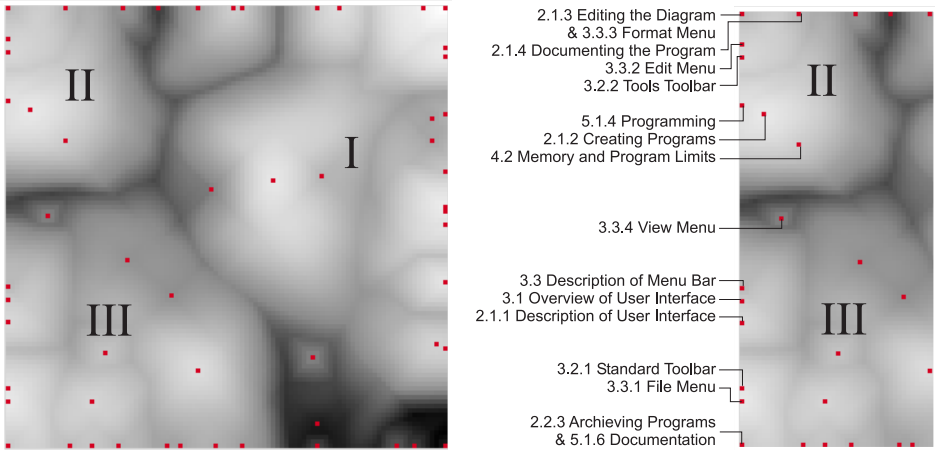

Abbildung 7: *Von links*: a) Inhaltliche Kapitelstrukturen. b) Beschriftung durch Kapitelüberschriften.

Wir entwickeln derzeit das System DocMINER (*Document Maps for INformation Elicitation and Retrieval*) für das Management wissenschaftlicher und technischer Dokumente. Hier setzt die Visualisierung der $\mathcal{P}^M$-Matrizen auf die Ergebnisse einer Vorverarbeitung mit Methoden des Information Retrieval und einer Dimensionsreduktion auf [4, 3]. Erfahrungen mit ersten Praxisanwendungen sind vielversprechend. So wurde das System bereits erfolgreich zur Untersuchung von Anwendungsszenarien für die Spezifikation von Systemkomponenten [2, 14] und für die Qualitätsüberprüfung von Be-

nutzerdokumentationen in Hinblick auf Struktur oder Mehrfachbeschreibung von Inhalten sowie zur Identifikation sogenannter *Single Sources* (einheitliche betriebliche Informationsquellen) eingesetzt [1]. Die Größe interessanter Dokumentensammlungen reicht dabei von rund hundert (Strukturanalyse einzelner Handbücher) bis zu einigen tausend Dokumenten variabler Länge (z.B. Sammlung von *Request for Comments*, das sind Definitionen der Protokolle und Verfahrensweisen des Internet).

Abbildung 7a zeigt die Dokumentenlandkarte eines Benutzerhandbuchs, das ein Schaltgerät zur Automatisierung von Gerätesteuerungen und seine Programmierung mit Hilfe eines Windows-Programms beschreibt. Jeder Punkt auf der Karte stellt ein einzelnes Kapitel dar. Insgesamt ergibt sich ein Bild der inhaltlichen Zusammenhänge der einzelnen Abschnitte: Gebiet I enthält im wesentlichen die Bibliothek der Schaltgerätfunktionen, Bereich II faßt Informationen zu dessen Programmierung zusammen, und Gebiet III beinhaltet Dokumente zur Bedienungsoberfläche, die sich mit der Simulation, Dokumentation und Übertragung des erstellten Programms vom PC auf das Schaltgerät befassen.

Eingesetzt als Benutzerschnittstelle für die „Navigation“ in Online-Dokumentationen hilft die Karte, inhaltlich verwandte Bereiche zu erkunden. So beschreibt etwa *Kapitel 2.1.2* (vgl. Abbildung 7b, oben links) im Rahmen eines Tutorials die Schritte zur Erstellung eines Programms, während *Kapitel 5.1.4* ergänzend die Programmierung einer Beispielapplikation darstellt.

Das System wird derzeit u.a. um eine „Sichtenkomponente“ erweitert, die es ermöglicht, inhaltliche Verbindungen von Dokumenten auf der Grundlage individueller Interessen zu erkennen - eine für das Wissensmanagement bedeutende Funktionalität.

6 Verwandte Arbeiten

Erregungsmuster einzelner Objekte leisten eine Dimensions- aber keine Datenreduktion. Insofern erinnern sie an ikonische oder geometrische Visualisierungstechniken (s. [15] für eine Übersicht). Hierzu gehören etwa die *Chernoff-Faces* [6] als wohl bekannteste Vertreter einer icon-basierten Technik und die *Technik der Parallelen Koordinaten* von *Inselberg* [12, 13] als Beispiel für eine Projektionstechnik, die u.a. im Data Mining-System WINVIZ [23, 22] implementiert ist. Gemeinsam ist diesen Techniken, daß jeder Merkmalsvektor separat und anhand seiner Merkmale nach festen Vorschriften in ein grafisches Abbild gebracht wird. Ähnlichkeiten der Merkmalsvektoren spiegeln sich dann in den Bilder wider, die somit für eine visuelle Objektgruppierung genutzt werden können.

Dies gleicht der Visualisierung unter Verwendung von Erregungsmustern. In beiden Fällen bleibt die Anzahl der noch sinnvoll zu analysierenden Merkmalsvektoren stark begrenzt, auch wenn aufgrund anderer Darstellungsformen bei den oben genannten Techniken zumeist die gleichzeitige Darstellung einer größeren Anzahl von Visualisierungen möglich ist. Demgegenüber erfolgt die Berechnung der Erregungsmuster unabhängig von der Anzahl der Merkmale. In der Berechnung liegt auch der wesentliche Unterschied. So gehen in jedes Erregungsmuster implizit alle und nicht nur ein einzelner Merkmalsvektor ein. Gegenüber den angesprochenen Techniken wird damit eine sinnvolle Überlagerung von Mustern und somit eine Datenreduktion möglich.

Die Nutzung von Merkmalskarten zur visuellen Datenreduktion ist keine neue Anwendung von Kohonens Modell (vgl. etwa [29], [30], [19], [36], [11], ...). Die meisten Arbeiten nutzen dabei die Verteilung der Zustandsvektoren, um eine Darstellung ähnlich der $\mathcal{P}^{M=X}$-Matrix zu gewinnen. An unserem Beispiel haben wir illustriert, daß diese Verteilung die Dichte der Merkmalsvektoren im Merkmalsraum reflektiert. Demnach sind Zustandsvektoren in Teilräumen einander ähnlicher, in denen Punkthaufen der Merkmalsvektoren liegen, während der leere Raum dazwischen vergleichsweise spärlich ausgefüllt ist. Folglich können Strukturen über einen paarweisen Distanzvergleich von Zustandsvektoren benachbarter Ausgabeeinheiten nach dem (wohl ersten) Vorschlag von *Ultsch* und *Siemon* [29] sichtbar gemacht werden. Hierzu werden die Abstände in den sog. *unified distance matrices*

(U-Matrizen) geordnet und gespeichert und anschließend über Graustufenbilder visualisiert. Die Distanzen eines Zustandsvektors werden dabei oft zu einem einzigen Wert zusammengefaßt (etwa dem Mittelwert [19, 11]), um die Matrizengröße an die Größe des verwendeten Gitters anzugleichen.

Im visuellen Bildeindruck können die resultierenden Graustufenbilder dann einer Überlagerung aller Erregungsmuster gleichen. Unterschiede ergeben sich jedoch aus der Berechnung. In den Bildern der $\mathcal{P}^M$-Matrizen erfolgt diese unabhängig von der Gesamtverteilung der Zustandsvektoren im Merkmalsraum. Wir sehen darin zwei Vorteile. Zum einen ist mit jedem Bildpunkt eine Bedeutung in Bezug auf die Merkmalsvektoren verbunden, nämlich die Abweichung eines bestimmten Zustandsvektors zu einem bestimmten Merkmalsvektor. Zum anderen werden Strukturen auch unabhängig von der erreichten Generalisierung in dem Sinne sichtbar, daß auch dann gute Visualisierungsergebnisse zu erwarten sind, wenn die Zustandsvektoren den gesamten Merkmalsraum approximieren und nicht überwiegend auf eine Wiedergabe der Merkmalsvektoren beschränkt bleiben. Beide Aspekte gemeinsam bilden die Basis der maschinellen MIDAS-Regelgenerierung und Anwendung sowie der interaktiven Analyseoperationen auf den Gaustufenbildern.

Im schrittweisen Aufbau einer Folge von $\mathcal{P}^M$-Matrizen sehen wir zudem eine bislang noch nicht beschriebene Form der Interaktion, die einem Analysten die Möglichkeit bietet, quasi „visuell“ durch den abgebildeten Merkmalsraum zu navigieren.

7 Zusammenfassung

Mit diesem Beitrag haben wir gezeigt, daß auch in der Kombination von wenigen Techniken aus verschiedenen Bereichen des Data Mining bislang kaum beachtete Potentiale liegen. Das MIDAS-System verfolgt hierzu einen explorativen Data Mining-Ansatz, der Visualisierungtechniken zur Strukturerkennung durch neuronale Merkmalskarten über eine Fuzzy-Term-Generierung mit maschinellen Lernverfahren zur Strukturbeschreibung kombiniert.

In diesem Beitrag haben wir uns auf eine Strukturerkennung beschränkt und mit den visuellen Darstellungen der Merkmalsverteilungen und den $\mathcal{P}^M$-Matrizen zwei Basis-Techniken der MIDAS-Datenvisualisierung vorgestellt. Diese beziehen den Benutzer sowohl in die Erkennung der Strukturen als auch in ihre Beschreibung ein. So entsteht ein Brückenschlag zwischen menschlicher Intuition und Sachkompetenz einerseits und datengetriebenen maschinellen Techniken andererseits.

Unsere Anwendungserfahrungen zeigen, daß gerade diese enge Kopplung dem Benutzer für eine fruchtbare Datenanalyse entgegenkommt. Die Visualisierungstechniken sind von verschiedenen Anwenderkreisen, u.a. Analysten aus dem Einzelhandel und technischen Redakteuren, gut verstanden und nutzbringend eingesetzt worden. Eine empirische Untersuchung der Benutzerakzeptanz und der Verwendbarkeit ist allerdings zur entgültigen Bewertung noch sinnvoll und auch notwendig.

Literatur

[1] A. Becks and M. Host. Qualitätsprüfung mit Dokumentenlandkarten. In *tekom-Frühjahrstagung – Gesellschaft für technische Kommunikation e.V.*, 1999. Stuttgart.

[2] A. Becks and J. Köller. Automatically Structuring Requirements Scenarios. In *Proc. of the 14th IEEE Intern. Conference on Automated Software Engineering*, October 1999. Cocoa Beach, Florida. To appear.

[3] A. Becks, S. Sklorz, and M. Jarke. Document maps: Semantic structuring of technical document collections. Report Series of the ESPRIT Project CREWS, Report No. Crews-99-05, RWTH Aachen, Germany, 1999. http://sunsite.informatik.rwth-aachen.de/CREWS/.

[4] A. Becks, S. Sklorz, and C. Tresp. Semantic Structuring and Visual Querying of Document Abstracts in Digital Libraries. In *Proc. of the 2nd European Conference on Research and Advanced Technology for Digital Libraries*, volume LNCS 1513. Springer, Berlin, September 1998. Crete, Greece.

[5] H. Chen, C. Schuffels, and R. Orwig. Internet Categorization and Search: A Self-Organizing Approach. *Journal of Visual Communication and Image Representation*, 7(1):88–102, 1996.

[6] H. Chernoff. The Use of Faces to Represent Points in k-Dimensional Space Graphically. *Journal Amer. Satistical Association*, 68:361–368, 1973.

[7] SPSS Clementine Data Mining System. Integral Solutions Limited, WWW http://www.isl.co.uk/, 1999.

[8] DataEngine. Management Intelligenter Technologien GmbH, WWW http://www.mitgmbh.de/, 1999.

[9] G. Deichsel and H. J. Trampisch. *Clusteranalyse und Diskriminanzanalyse*. Gustav Fischer Verlag Stuttgart, 1985.

[10] M. Gebhardt, M. Jarke, M. A. Jeusfeld, C. Quix, and S. Sklorz. Tools for data warehouse quality. In *Proc. of the 10th Intern. Conf. on Scientific and Statistical Database Management (SSDBM'98)*, pages 229 – 232. IEEE CS Press, July 1998. Capri, Italy.

[11] J. Iivarinen, T. Kohonen, J. Kangas, and S. Kaski. Visualizing the clusters on the self-organizing map. In *Proc. of the Conf. on Artificial Intelligence Research in Finnland*, pages 122–126. Finnish Artificial Intelligence Society, Helsinki, Finland, 1994.

[12] A. Inselberg. The Plane with Parallel Coordinates, Special Issue on Computational Geometry. *The Visual Computer*, 1:69–97, 1985.

[13] A. Inselberg and B. Dimsdale. Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry. In *Proc. of Visualization'90*, pages 361–370, 1990. San Francisco, CA.

[14] M. Jarke, A. Becks, J. Köller, C. Tresp, and B. Braunschweig. Designing Standards for Open Simulation Enviroments in Chemical Industries: A Computer-Suppoted Use-Case Approach. In *Proc. of INCOSE'99, Systems Engineering – Sharing the Future*, page to appear, June 1999. Brighton, GB.

[15] D. A. Keim and H.-P. Kriegel. Visualization Techniques for Mining Large Databases: A comparison. *IEEE Transaction on Knowledge and Data Engineering*, 8(6):923–938, December 1996.

[16] Data Mining mit KEPLER. Dialogis Software & Services GmbH, WWW http://www.dialogis.de/, 1999.

[17] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

[18] T. Kohonen. *Self-Organizing Maps*. Springer Verlag, second edition edition, 1997.

[19] M. A. Kraaijveld, J. Mao, and A. K. Jain. A non-linear projection method based on Kohonen's topology preserving maps. In *Proc. of the 11th IEEE Intern. Conf. on Pattern Recognition*, pages 41–45. IEEE Computer Society Press, CA, 1992.

[20] K. Lagus. Generalizability of the WEBSOM Method to Document Collections of Various Types. In *Proc. of the 6th European Congress on Intelligent Techniques and Soft Computing (EUFIT'98)*, volume 1, pages 210–214, 1998. Aachen, Germany.

[21] K. Lagus, T. Honkela, S. Kaski, and T. Kohonen. Self-Organizing Maps of Document Collections: A New Approach to Interactive Exploration. In *Proc. of the Second Intern. Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 238–243. AAAI Press, Menlo Park CA, August 1996. Portland, Oregon.

[22] H.-Y. Lee and H.-L. Ong. Visualization Support for Data Mining. *IEEE Expert Journal*, 11(5):69–75, October 1996.

[23] H.-Y. Lee, H.-L. Ong, and L.-H. Quek. Exploiting Visualization in Knowledge Discovery. In *Proc. of the First Intern. Conf. on Knowledge Discovery and Data Mining (KDD'95)*, pages 198–203, 1995.

[24] X. Lin, D. Soergel, and G. Marchionini. A Self-Organizing Semantic Map for Information Retrieval. In *Proc. of the 14th Annual Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 262–269, 1991.

[25] O. L. Mangasarian and W. H. Wolberg. Cancer diagnosis via linear programming. *SIAM News*, 23(5):1–18, September 1990.

[26] H. Ritter. *Selbstorganisierende neuronale Karten*. PhD thesis, TU München, Theoretische Physik, Germany, 1988.

[27] S. Sklorz. A method for data analysis based on self-organizing feature maps. In *Proc. of WAC'96, Soft Computing with Industrial Applications*, volume 5, pages 611–616. TSI Press Series, Albuquerque, NM, May 1996. Montpellier, France.

[28] S. Sklorz and M. Mücke. A hybrid approach for medical data analysis. In *Proc. of the 5th European Congr. on Intelligent Techniques and Soft Computing (EUFIT'97)*, volume 2, pages 1162–1166, Aachen, Germany, September 1997.

[29] A. Ultsch and H. P. Siemon. Exploratory Data Analysis: Using Kohonen Networks on Transputers. Technical Report No. 329, Fachbereich Informatik, Universität Dortmund, Germany, 1989.

[30] A. Ultsch and H. P. Siemon. Kohonen's self-organizing feature maps for exploratory data analysis. In *Proc. of ICNN'90, Intern. Neural Network Conference*, pages 305–308. Kluwer, Dortrecht, 1990.

[31] T. Villmann. *Topologieerhaltung in selbstorganisierenden neuronalen Merkmalskarten*. Verlag Harri Deutsch, Frankfurt, 1996. PhD thesis, Univ. Leipzig, Germany.

[32] R. Weber and H.-J. Zimmermann. Automatische Akquisition von unscharfem Expertenwissen. *Fachbeiträge der Zeitschrift KI*, pages 20–26, 1991.

[33] W. H. Wolberg and O. L. Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In *Proc. of the National Academy of Science, U.S.A.*, volume 87, pages 9193–9196, December 1990.

[34] S. Wrobel, D. Wettschereck, E. Sommer, and W. Emde. Extensibility in data mining systems. In *Proc. of the 2nd Intern. Conf. on Knowledge Discovery and Data Mining (KDD'96)*, pages 214–219. AAAI Press, CA, August 1996. Oregon, USA.

[35] J. Zhang. Selecting typical instances in instance-based learning. In *Proc. of the Ninth Intern. Machine Learning Conference*, pages 470–479. Morgan Kaufmann, 1992.

[36] X. Zhang and Y. Li. Self-organizing map as a new method for clustering and data analysis. In *Proc. of the IEEE Intern. Joint Conf. on Neural Networks*, pages 2448–2451, 1993. Nagoya.