Datenbionik: Selbstorganisierende Systeme zur Entdeckung ungewöhnlicher Strukturen in Unternehmensdaten

Alfred Ultsch,
Lehrstuhl Datenbionik, FB12,
Universität Marburg, Hans Meerwein Str. 22,
35032 Marburg,
ultsch@ulweb.de

Was ist Datenbionik?

Bionik kann definiert werden als das „Studium der biologischen Evolution aus der Sicht des Ingenieurs“ [Rechenberg13]. Die Idee dabei ist, dass die Evolution als ein universelles Optimierungsverfahren angesehen werden kann, welches für mache Probleme eine ideale Lösung entwickelt hat. Im Fokus der Bionik stehen heutzutage vor allem Mechanik und Werkstofftechnik, wie z.B. der sog. Lotuseffekt. Die Datenbionik hingegen untersucht speziell die Informationsverarbeitung in der Natur. Sie versucht die Prinzipien und Methoden, die hierfür in der Natur erkannt werden können, in den Computer zu übertragen. Ein klassisches Beispiel ist die Übertragung der Funktionsweise von biologischen Neuronalen Netzwerken, wie sie z.B. in Gehirnen realisiert sind, in künstliche neuronale Netze [Hebb 49], [WikiANN13].

In der Natur lassen sich Informationsverarbeitungsalgorithmen aber nicht nur in Neuronalen Netzwerken entdecken. In der Genetik z.B. wird die DNA als Universalspeicher sämtlicher Informationen angesehen, die für den Aufbau, die Funktionsweise und die Vererbung notwendig sind. Aus der Sicht der Datenbionik interessieren dabei vor allem die Algorithmen zur Datensicherheit (Redundanz, Fehlerkorrektur, Selbstverdoppelung) und die der Optimierung durch evolutionäre Algorithmen [Gerdes et al 13].

Schwärme von Vögeln, Bienen, Ameisen und manchmal auch Menschen, zeigen ein kollektives Verhalten, welches von lästig (Stau auf Autobahnen) bis zu hochinteressant (Panikverhalten) variiert. Das interessanteste Phänomen, welches in Schwärmen beobachtet werden kann, ist die Emergenz. Emergenz bedeutet das Auftauchen von makroskopischen Strukturen welche sich nicht einfach auf das (mikroskopische) Verhalten einzelner Mitglieder des Schwarms zurückführen lässt (siehe Bild 2).

Wo ist das Problem?

Datenverarbeitung, so wie sie gegenwärtig in unseren EDV-Systemen praktiziert wird, hat als Grundprinzip den EVA-Algorithmus: **E**ingabe-**V**erarbeitung-**A**usgabe (siehe Bild 1).

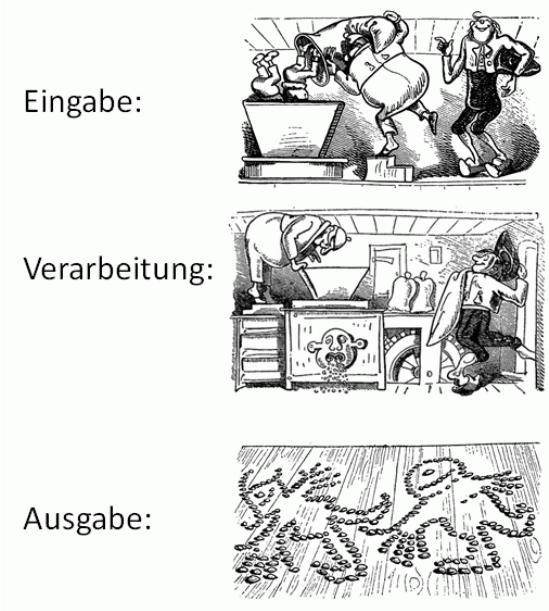

Bild 1: Wilhelm Busch's EVA-Algorithmus bei Max und Moritz

Dieser Algorithmus hat sich in vielen praktischen Anwendungen als nützlich erwiesen. Die Telekom hat z.B. in Magdeburg ein enorm leistungsfähiges Rechenzentrum für IT-Dienstleistungen aufgebaut [DataCenter10]. Solche EDV-Zentren ermöglichen es, dass rechtzeitig zum Monatsende für 10^9 und mehr Kunden eine Abrechnung erstellt wird, die jedes einzelne Telefongespräch berücksichtigt. Unter der Annahme, jeder Kunde führt im Monat im Mittel 100 Telefonate, ist der Aufwand für eine solche Berechnung in der Größenordnung O (10^{11}). Das bedeutet, dass mindestens 100.000.000.000 Rechenoperationen rechtzeitig für die Rechnungsstellung bewältigt werden müssen.

Andere Fragestellungen auf solchen, heute bereits bestehenden, Datensammlungen sind jedoch noch komplexer. In Zeiten umkämpfter Marktanteile ist es für ein Unternehmen enorm wichtig zu wissen, warum Kunden den Service kündigen und eventuell zur Konkurrenz wechseln. Solche Kunden werden im Fachjargon „Churner" (Change and Turn) genannt. Nehmen wir an, ein TelCom Unternehmen speichert für jeden seiner Kunden ca. 50 Merkmale, die hierfür wichtig sein könnten. Dazu gehören neben den Verbindungsdaten wie Anzahl, Tages-, Wochen-, Feiertagsnutzung und Netzwerknutzung auch soziodemographische Daten, wie Alter, Geschlecht, Wohnort, Zahlungsverhalten, Bildungsstand etc. Jede beliebige Kombination dieser angenommenen 50 Merkmale (alias Dimensionen, Variablen, Komponenten) könnte die Churner beschreiben. Wenn im Mittel nur 2 Ausprägungen (z.B. männlich/weiblich, wenig/viel) für die Merkmale angenommen werden ergibt

ein Problem in der Größenordnung von $2^{50} > 10^{15}$ für jeden der angenommenen 10^9 Kunden.

Fragestellungen wie z.B. „Gibt es Gruppen von Churnern, welche gemeinsame Verhaltensweisen besitzen?“ sind somit mit heutigen EVA-Methoden sehr schlecht zu behandeln. Diese setzen voraus, dass in den Programmen bzw. Datenbankanfragen bereits festgelegt ist, wonach genau zu suchen ist. Eine auch nur annähernd vollständige Suche in einem Suchraum mit der oben skizzierten Größe ist heute praktisch nicht durchführbar.

Emergenz und Selbstorganisation

Bei der Suche nach neuen, bislang unentdeckten Strukturen in Datensammlungen kann ein Phänomen nützlich werden, welches in Natur und Biologie beobachtet werden kann. In Systemen, die aus vielen einzelnen mehr oder weniger gleichartigen Teilchen auf der mikroskopischen Ebene bestehen (Vielteilchen-System) kann gelegentlich das Auftauchen von makroskopischen Strukturen beobachtet werden. Ein Beispiel ist die La-Ola Welle (Mexican Wave) bei Sportveranstaltungen. Eine andere Beobachtung (Karstberger und Mitarbeiter) zeigte, dass sich Bienen ihrer Feinde erwehren, indem sie sich auf eine ebene Oberfläche setzen und dort eine große, die Feinde abschreckende Spirale erzeugen und entsprechend mit den Flügeln schlagen [Karstberger et al 13]. Siehe das beeindruckende Viedeo hierzu in **[BeeWave08].**

Bild 2: Emergenz einer Spirale gebildet aus Bienenflügeln, aus [Karstberger et al 13] Fig.3a

Das unerwartete und auch nicht vorab berechenbare Auftreten solcher Strukturen wird als Emergenz bezeichnet [Stephan 05]. Emergenz lässt sich nicht nur in biologischen Vielteilchen-Systemen beobachten, sondern auch in der unbelebten Natur: Füllt man in seiner Küche eine Pfanne mit Olivenöl stellt die Herdplatte auf die

höchste Hitzestufe, so kann man die im Bild 3 gezeigte Struktur beobachten. Es handelt sich dabei um sog. Bénard-Zellen, welche durch Konvektion, also die Auf- und Abwärtsbewegung des Olivenöls entsteht.

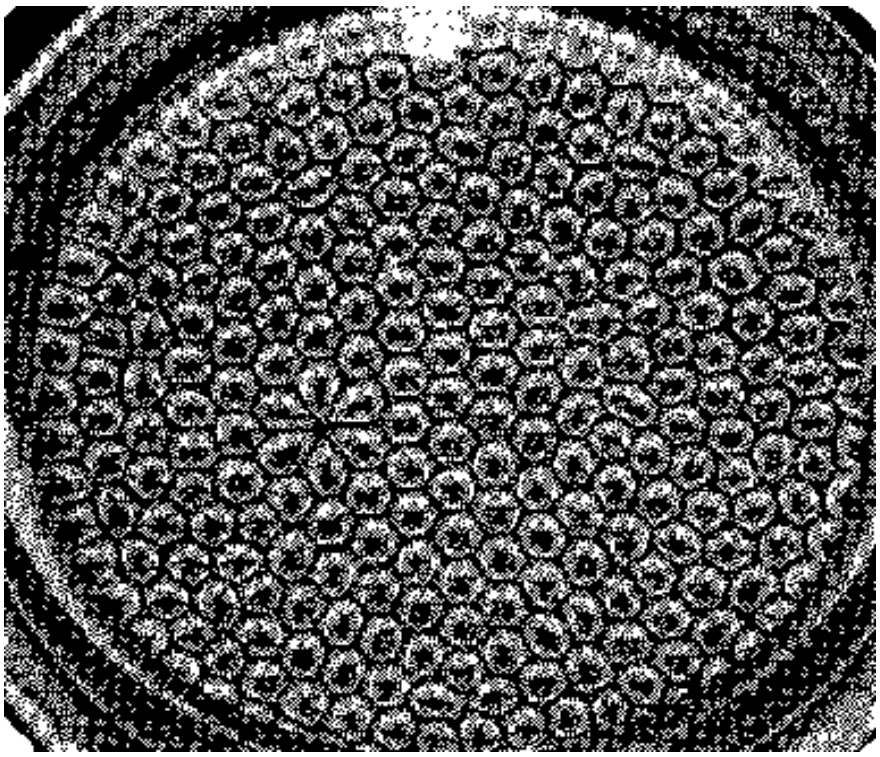

Bild 3: Emergenz in der häuslichen Bratpfanne (Bildquelle unbekannt)

Angetrieben wird die Konvergenz durch den Temperaturunterschied zwischen heißem Boden und kühlerer Oberfläche der Pfanne. Das Entstehen solcher emergenter Strukturen erfolgt i.d.R. ohne äußere steuernde Elemente (zentrale Kontrolle). Dies bezeichnet man als Selbstorganisation eines Systems. Bei den Konvektions-Zellen entsteht eine makroskopische Struktur in Gestalt der in Bild 3 zu sehenden Waben aus einer enormen Menge ($>10^{19}$ pro cm^3) von gleichartigen Ölmolekülen (Mikrostruktur).

Nutzbarmachung von Selbstorganisation und Emergenz im Computer

Der finnische Physiker Teuvo Kohonen hat 1982 ein künstliches Neuronales Netz von Neuronen, die sog. Selbstorganisierende Merkmalskarte (Self Organizing Map, SOM) entwickelt, welches die Selbstorganisationsprinzipien von Gehirnstrukturen nachbilden will [Kohonen 82]. Dies wurde 1990 um die U-matrix zur emergenten SOM (ESOM) ergänzt [Ultsch/Siemon 1990]. Hauptprinzip der ESOM ist die Abbildung der vielen Merkmale eines Datensatzes (hochdimensionaler Datenraum) auf eine (Land-)Karte in Form eines Gitters von Neuronen. Diese Abbildung (engl.: map) wird durch Selbstorganisation erzeugt. Jeder einzelne Datensatz sucht sich einen Platz auf der Karte und beeinflusst dabei seine Nachbarschaft. Insgesamt organisieren sich die Daten selbstständig auf der SOM-Karte, so dass einander

ähnliche Datensätze auch nahe beieinander auf der entstehenden Karte liegen. Einander unähnliche Datensätze können auch weit entfernt auf der Karte liegen. Bei der ESOM werden auch komplexe Verflechtungen innerhalb der hochdimensionalen Daten entflochten und die Daten nachbarschaftsgetreu (topologieerhaltend) dargestellt. Ein klassisches Beispiel hierfür ist der Chainlink Datensatz [FCPS 13]

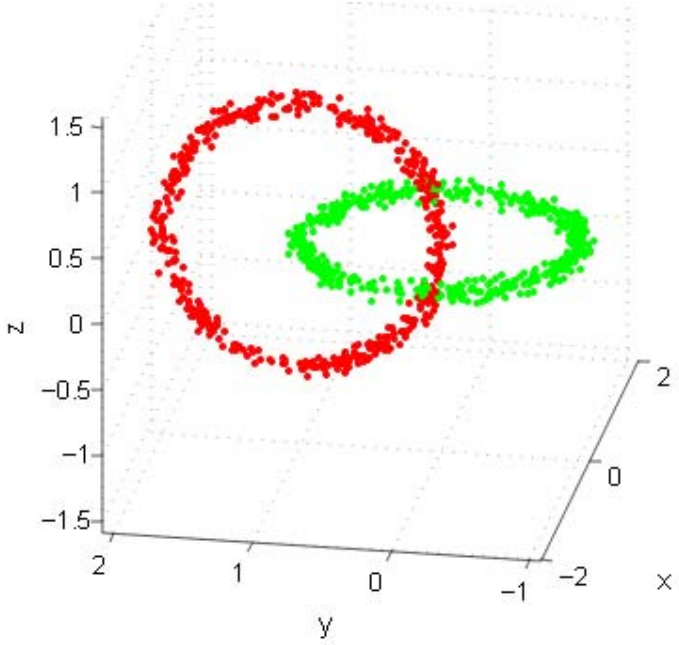

Bild 4: Chainlink Datensatz [FCPS 13]

Der Chainlink Datensatz besteht aus zwei klar getrennten Gruppen von Daten die jedoch im hochdimensionalen Raum (in diesem Fall dreidimensional) wie Kettenglieder ineinander verwoben sind. Kaum ein anderes Projektionsverfahren ist in der Lage, bei einer Abbildung auf eine 2-dimensionale Ebene (Karte) die beiden Kettenglieder als getrennt - jedoch als Gruppe in sich geschlossen (d.h. topologieerhaltend) - wiederzugeben. Die ESOM leistet diese Erhaltung solcher Nachbarschaftsbeziehungen, wie Bild 5 zeigt.

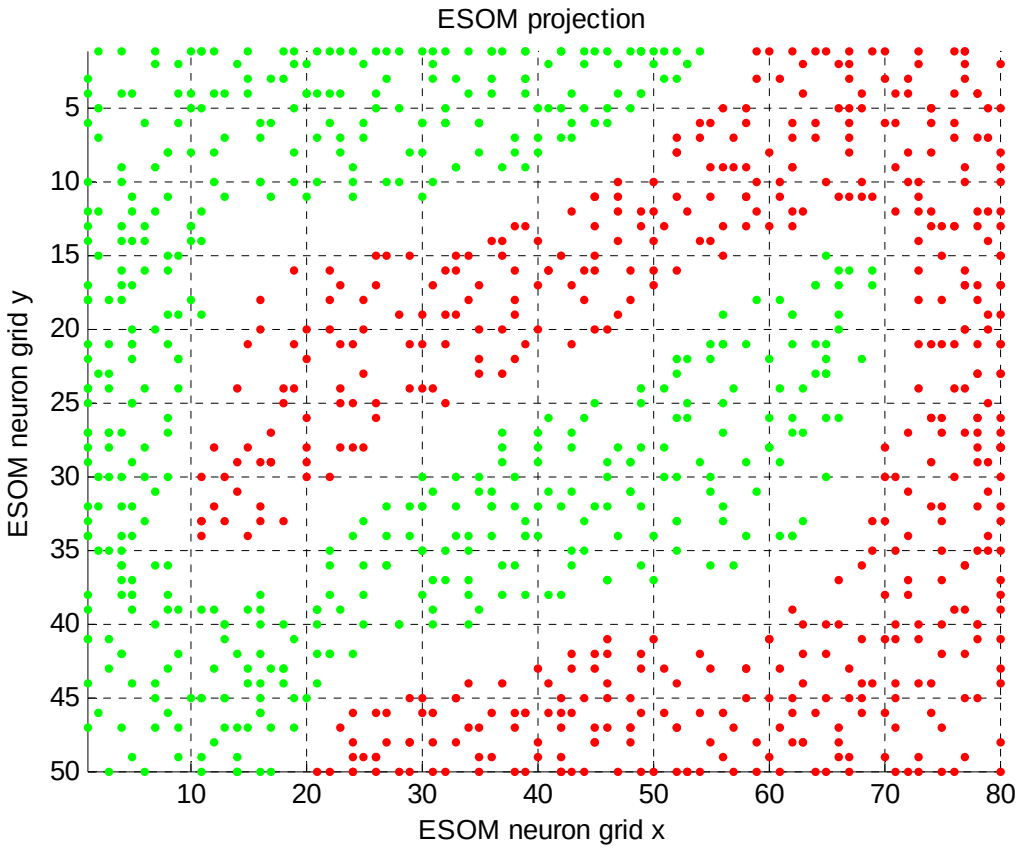

Bild 5: ESOM Projektion des Chainlink Datensatzes auf eine Landkarte mit 50x80 Neuronen

Bei der ESOM ist zu jedem Neuron auf dem Projektionsgitter der entsprechende Ort im Datenraum bekannt. Daher kann in der Nachbarschaft eines jeden Neurons eine mittlere Distanz der Daten im hochdimensionalen Datenraum bestimmt werden. Wird diese Distanz auf der Karte als Höhe abgebildet, so ergibt sich die U-Matrix [Ultsch 03]. Berge in einer U-Matrix bedeuten große Entfernungen im Datenraum. Liegen Daten gemeinsam in einem Tal in einer U-Matrix, so besitzen diese Daten starke Gemeinsamkeiten. Die Struktur der Täler, genauer gesagt die der Wasserscheiden, auf der entstandenen U-Matrix ist eine selbstorganisierende und emergente Struktur. Diese visualisiert die Distanzstrukturen eines komplexen Datensatzes in einer Form, die uns Menschen –von physischen Karten her geläufig ist.

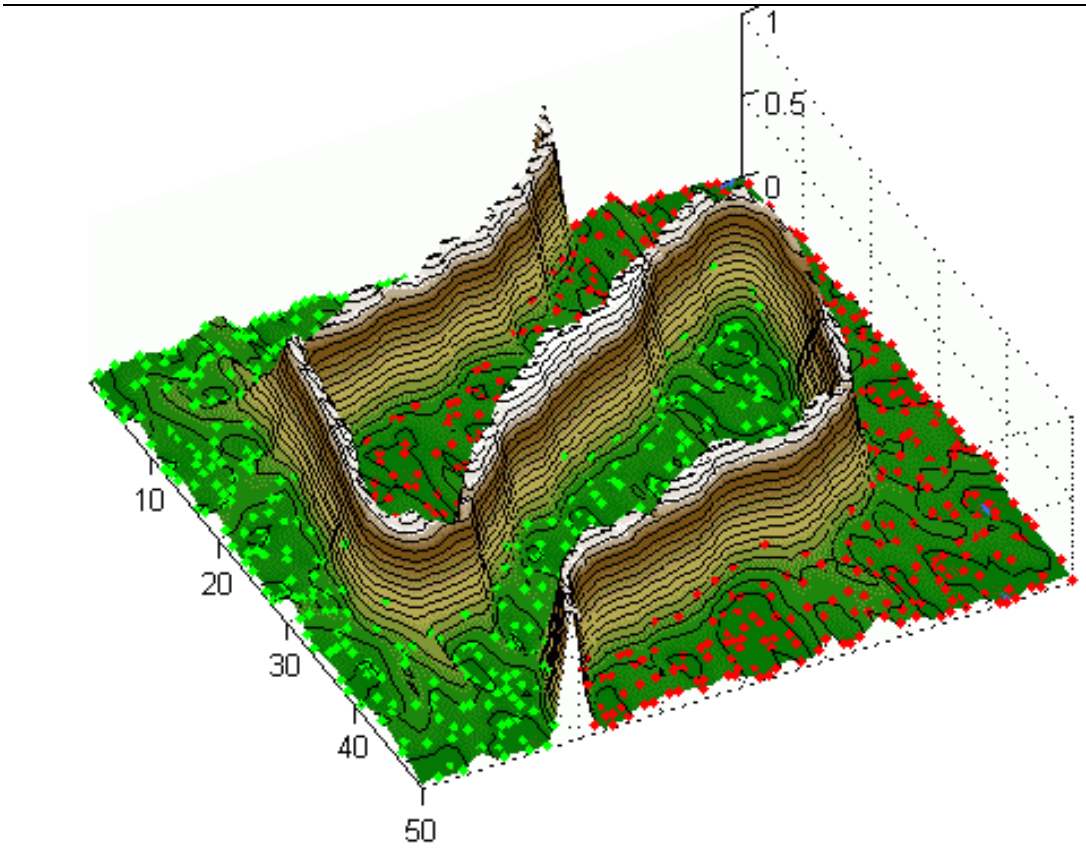

Bild 6: U-Matrix des Chainlink Datensatzes

Emergenz in Computerprogrammen zur Analyse von Daten

Zur Beantwortung der Fragestellung “Welche besonderen Strukturen finden sich in den Daten meines Unternehmens?“ kann die im letzten Kapitel beschriebene ESOM/U-Matrix-Methode angewendet werden. Dies nutzt die in der Natur beobachtbaren Phänomene Selbstorganisation und Emergenz. Hierfür gibt es bereits einige erfolgreiche Anwendungsbeispiele mit Datensammlungen aus den verschiedensten Bereichen [DataBio13]. In Zusammenarbeit mit der schweizer Firma Swisscom wurde z.B. eine solche Datenanalyse für Telekommunikationsdaten in Mobilfunknetzen durchgeführt. Für eine wissenschaftliche Veröffentlichung konnten wir eine kleine Teilmenge der zur Verfügung gestellten Daten verwenden [Ultsch 02]. Aus der sehr viel größeren Datensammlung wurden 21 Variablen ausgewählt, welche verschiedene Aspekte des Kundenverhaltens beschrieben: Zahlungsverhalten, Nutzung verschiedener Netzwerke, Ziele von Ferngesprächen, Nutzungszeiten, Nutzungsdauer der Netze etc. Die daraus entstandene U-Matrix zeigt Bild 6

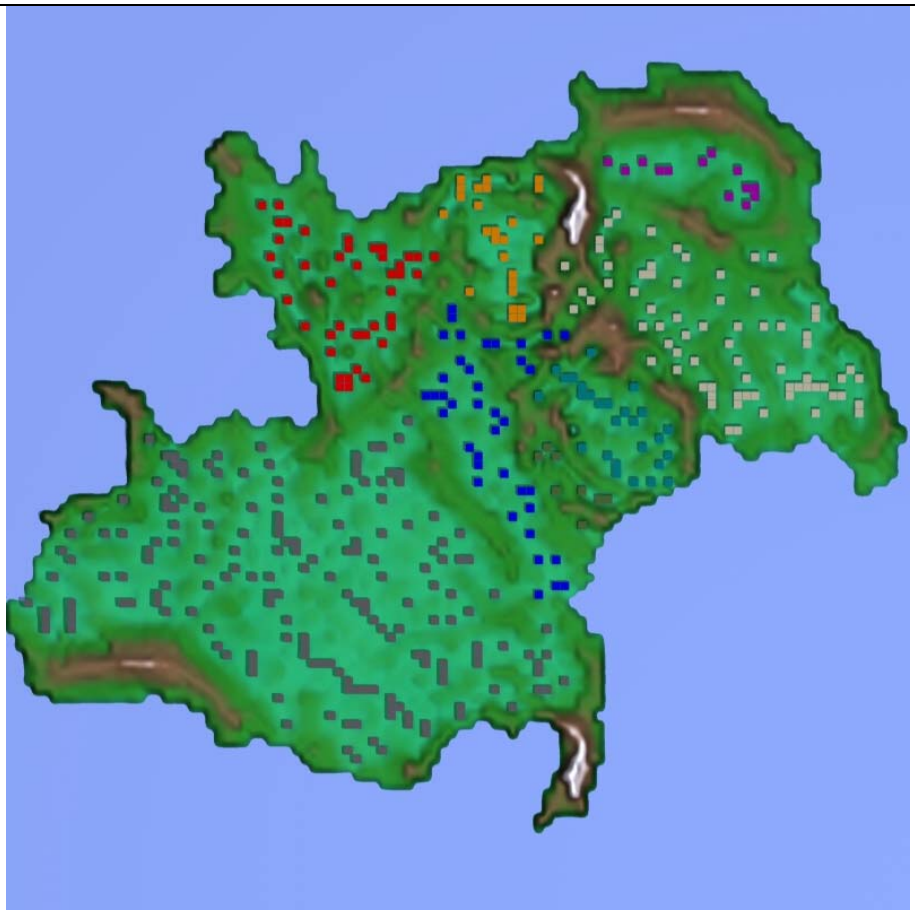

Bild 7: U-Matrix von Mobilfunk-Kundendaten

Diese U-Matrix ergab insgesamt 7 verschiedene Gruppen von Kunden, die sich durch ein für die jeweilige Gruppe typisches Nutzungsverhalten auszeichneten [Ultsch 02]. Dies wird durch die farbigen Markierungen in Bild 6 dargestellt.

Im Gegensatz zur ESOM/U-Matrix-Methode erfordern andere Verfahren zur Clusterung von hochdimensionalen Daten entweder die Vorgabe der Anzahl der Cluster, die zu bestimmen sind, oder sie besitzen eine implizite Annahme, wie die Cluster und ihre Begrenzungen strukturiert sein sollen. Die bekannte k-means Clustermethode beruht z.B. darauf, dass die Gestalt der Cluster immer kugelförmig sei und dass die Clustergrenzen durch Ebenen vorgegeben sind.

Mittels spezieller Verfahren des „Knowledge Discovery“ konnten in den Swisscom-Daten detailierte Gruppen identifiziert werden, für die es sich für das Unternehmen lohnte, die Abwanderungsquote zu senken. Ein auf menschliches Verstehen abzielendes Verfahren zur Beschreibung dessen, was die Besonderheit einer jeden Gruppe ausmacht, also die Regelgenerierung mit sig*, lieferte konkrete Hinweise darauf, weshalb die Kunden vermutlich abwandern [Ultsch 02]. Für die Swisscom erbrachte dies z.B. die Erkenntnis, dass eine lohnende Menge von Kunden (zahlungskräftig, hoher Umsatz), die von einer bestimmten Region in der Schweiz regelmäßig in ein bestimmtes EU-Land telefonieren, zum Abwandern neigen. Hier musste über die entsprechende Preisstruktur im Vergleich zu den Mittbewerbern nachgedacht werden.

Diskussion

Biologische Neuronale Netze wie z.B. Gehirne machen es offensichtlich möglich, dass sich sensorische Daten selbstständig zu bedeutungsvollen Einheiten, z.B. Objekten, Begriffen oder Konzepten, organisieren. Selbstorganisation bedeutet eine radikale Abkehr von der bislang geübten EVA-Praxis in der EDV. Bei Eingabe-Verarbeitung-Ausgabe (EVA) wird davon ausgegangen, dass die Daten eher so passiv sind wie das Mahlgut in Bild 1. Der Müller, d.h. der Programmierer, weiß dann, was damit zu tun ist. Selbstorganisation in der Datenverarbeitung bedeutet, dass Daten sich selbst sinnvoll verarbeiten. Daten finden sich zu zusammengehörigen Gruppen zusammen und grenzen sich selbstständig, ohne übergeordnete Kommandostruktur, von anderen ab.

Eine unterhaltsame Demonstration dieses Prinzips ist der von uns entwickelte MusikMiner [Mörchen et al 06]. Dieser erlaubt es, dass sich die auf einer Festplatte vorhandene Musik selbst organisiert. Die dabei entstehende U-Matrix liefert eine Gesamtschau der eigenen Musiksammlung und kann zur Entdeckung von Stilrichtungen im eigenen Musikgeschmack beitragen.

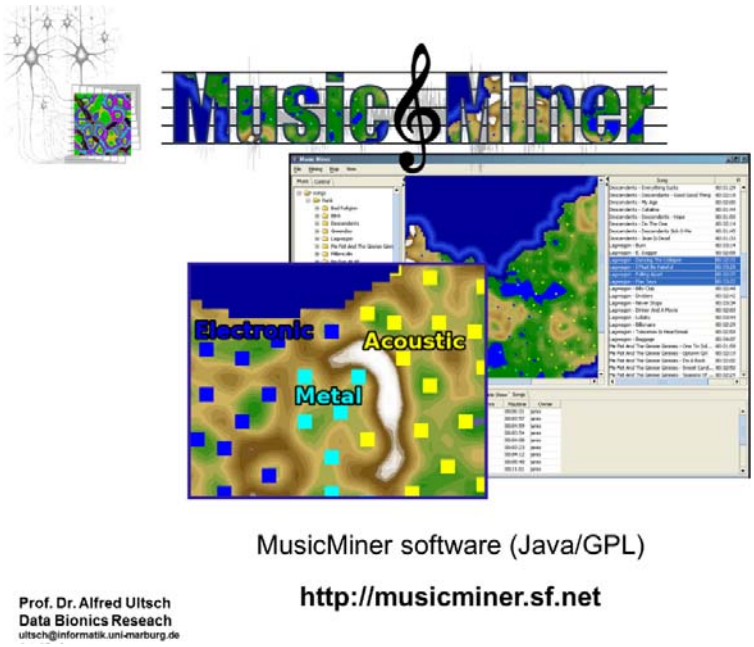

Bild 8: Selbstorganisation und Emergenz in persönlichen Musiksammlungen

Emergenz als Gestaltungsprinzip in der Natur ist in der Wissenschaft nicht unumstritten. Achim Stefans Buch liefert die zugehörige wissenschaftstheoretische Diskussion [Stephan 05]. Jedoch verspricht die Nutzung von Emergenz in der Datenverarbeitung die Entdeckung von absolut neuem, bislang nicht bekanntem Wissen in Datensammlungen, z.B. eines Unternehmens. Hierzu gehört auch die Entdeckung von sonderbarem, abweichendem Verhalten, wie z.B. Kreditkartenbetrug oder betriebliche Unregelmäßigkeiten.

Einem dabei ungerechtfertigten Gebrauch des Begriffs Emergenz kann entgegengetreten werden, wenn klare definiert wird, wann ein System der Emergenz fähig ist und woran man emergente Phänomene erkennen kann. Der Autor hat eine pragmatische Definition von Emergenz für die Nutzung in der Informatik vorgeschlagen [Ultsch 07]. Die wesentlichen Forderungen sind dabe erstens, dass das zugrundeliegende System ein Vielteilchen-System mit (zumindest lokaler) nichtlinearer Interaktion (dissipative Strukturen [Nicolis/Prigogine 77])[. Übersetzt: mit 10 Leuten in einem Stadion lässt sich kaum eine La-Ola Welle machen. Die zweite Anforderung für emergente Systeme in der Informatik ist, dass die entstehenden makroskopischen Strukturen sich nicht einfach auf die mikroskopischen Funktionen zurückführen lassen. Das heißt,: auch bei einer Obduktion aller Besucher eines Stadiums werden keinerlei Hinweise auf Entstehung, Amplitude, Frequenz, Wiederholrate etc. einer LaOla Welle gefunden.

Das bedeutet auch, dass, um in der Natur eine Wasserscheide, z.B. die europäische Hauptwasserscheide zwischen Donau und Rhein, erkennen zu können, es unzureichend ist, lokale Höhenlinien anzusehen. Hierzu muss die „Gestalt“ der gesamten Landschaft betrachtet werden.

Bild 9: Wasserscheiden in Deutschland [NordNordWest/Wikipedia]

Da die U-Matrix-Methode zur Erkennung von Gruppen in Daten genau auf der Erkennung von Wasserscheiden beruht, wie in einer von oben betrachteten Landschaft, kann diese Methode als ein erstes Beispiel für datenbionische Umsetzung der in der Natur beobachtbaren Prinzipien „Selbstorganisation“ und „Emergenz“ betrachtet werden.

Zusammenfassung

Die heutigen Verfahren der EDV beruhen i.d.R. darauf, dass Daten als passiv angesehen werden. Aktiv sind Programme, in denen das Wissen, was mit den Daten zu tun ist, eingebaut wurde. Das Ziel der Verarbeitung muss also vorab bekannt sein. Diese Vorgehensweise ist erfolgreich bei Standardanwendungen der Informatik wie z.B. Buchhaltung, Verwaltung von Kundendaten und effiziente Speicherung großer Datenmengen in (Cloud-) Datenbanken.

Für offene Fragestellungen wie z.B. *„Gibt es in meinen Daten etwas Besonderes?“* könnten sich dagegen datenbionische Methoden als nützlich erweisen. Diese übertragen Prinzipien der Informationsverarbeitung aus der Natur in Computerprogramme. Gerade für die Suche nach neuartigem, bisher unerkanntem und nützlichem Wissen in gegebenen Datensammlungen bieten sich die datenbionische Methoden der Selbstorganisation und Emergenz an. Bestimmte künstliche Neuronale Netzen erlauben bereits heute die Selbstorganisation von Daten. Auf den dabei entstehenden dreidimensionalen Landschaften können emergente Strukturen mit Hilfe einer U-Matrix erkannt werden. Erste Anwendungen auf Firmendaten demonstrieren die Anwendbarkeit dieser Methode für die Praxis.

Emergenz und Selbstorganisation stehen erst am Anfang einer spannenden Neuausrichtung der Methoden und Grundprinzipien in der Verarbeitung von größeren Datenmengen.

Literatur

[Hebb 49] Hebb,D: The organization of behavior. A neuropsychological theory. Erlbaum Books, Mahwah, N.J. 2002, Nachdruck der Ausgabe, New York ,1949.

[Gerdes et al 13] Gerdes,I., Klawonn,F., KruseR.: Evolutionäre Algorithmen,Vieweg, Wiesbaden 2004

[Karstberger et al 13] Kastberger G, Weihmann F, Hoetzl T.: Social waves in giant honeybees (Apis dorsata) elicit nest vibrations, Naturwissenschaften. 2013 Jul;100(7):595-609

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

[Mörchen et al 06] Moerchen, F., Ultsch, Noecker, M., Stamm, C.: Databionic visualization of music collections according to perceptual distance, Proceedings 6th International Conference on Music Information Retrieval London, UK, pp. 396-403, 2005.

[Mörchen/Ultsch 05] Ultsch, A., Moerchen, F.: ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM, Technical Report Dept. of Mathematics and Computer Science, University of Marburg, Germany, No. 46, (2005).

[Nicolis/Prigogine 77] Nicolis, G. and Prigogine, I.: Self-Organization in Nonequilibrium Systems, Wiley-Interscience, New York, 1977.

[Stephan 05] Stephan, A.: Emergenz, Mentis, 2005

[Ultsch 02] Ultsch, A.: Emergent Self-Organizing Feature Maps used for Prediction and Prevention in Mobile Phone Markets, Jornal of Targeting 10/4, Steward, London (2002), pp. 401-425

[Ultsch 03] Ultsch, A.: Maps for the Visualization of high-dimensional Data Spaces, In Proceedings Workshop on Self-Organizing Maps (WSOM 2003), Kyushu, Japan, (2003), pp. 225-230.

[Ultsch 07] Ultsch, A.: Emergence in Self-Organizing Feature Maps, In Proceedings Workshop on Self-Organizing Maps (WSOM '07), Bielefeld, Germany, 2007

[Ultsch/Siemon 90] Ultsch, A. , Siemon, H. P.: Kohonen's self organizing feature maps for exploratory data analysis. In Proceedings of ICNN'90, International Neural Network Conference, pages 305-308, Kluwer, (1990)

Weblinks

[BeeWave08] http://news.nationalgeographic.com/news/2008/09/080912-bee-wave.html

[DataBio13] http://www.uni-marburg.de/fb12/datenbionik

[DataCenter10] http://www.datacenterinsider.de/themenbereiche/cloud/colocation/articles/269910

[FCPS13] http://www.uni-marburg.de/fb12/datenbionik/data

[Rechenberg13] http://www.bionik.tu-berlin.de/

[Telekom10] http://www.datacenter-insider.de/themenbereiche/cloud/colocation/articles/269910

[WikiANN13] http://en.wikipedia.org/wiki/Artificial_neural_network