Wissuwa, Stefan; Cleve, Jürgen; Lämmel, Uwe

Working Paper

Analyse zeitabhängiger Daten durch Data-Mining-Verfahren

Wismarer Diskussionspapiere, No. 21/2005

Provided in cooperation with:
Hochschule Wismar

Hochschule Wismar
University of Technology, Business and Design
Fachbereich Wirtschaft

Hochschule Wismar
University of Technology, Business and Design
Faculty of Business

Stefan Wissuwa, Jürgen Cleve, Uwe Lämmel

Analyse zeitabhängiger Daten durch Data-Mining-Verfahren

Heft 21/2005

W D P

Wismarer Diskussionspapiere / Wismar Discussion Papers

Der Fachbereich Wirtschaft der Hochschule Wismar, University of Technology, Business and Design bietet die Präsenzstudiengänge Betriebswirtschaft, Management sozialer Dienstleistungen, Wirtschaftsinformatik und Wirtschaftsrecht sowie die Fernstudiengänge Betriebswirtschaft, International Management, Krankenhaus-Management und Wirtschaftsinformatik an. Gegenstand der Ausbildung sind die verschiedenen Aspekte des Wirtschaftens in der Unternehmung, der modernen Verwaltungstätigkeit im sozialen Bereich, der Verbindung von angewandter Informatik und Wirtschaftswissenschaften sowie des Rechts im Bereich der Wirtschaft.
Nähere Informationen zu Studienangebot, Forschung und Ansprechpartnern finden Sie auf unserer Homepage im World Wide Web (WWW): http://www.wi.hs-wismar.de/.

Herausgeber: Prof. Dr. Jost W. Kramer
Fachbereich Wirtschaft
Hochschule Wismar
University of Technology, Business and Design
Philipp-Müller-Straße
Postfach 12 10
D – 23966 Wismar
Telefon: ++49/(0)3841/753 441
Fax: ++49/(0)3841/753 131
e-mail: j.kramer@wi.hs-wismar.de

Vertrieb: HWS-Hochschule Wismar Service GmbH
Phillipp-Müller-Straße
Postfach 12 10
23952 Wismar
Telefon:++49/(0)3841/753-574
Fax: ++49/(0)3841/753-575
e-mail: info@hws-startupfuture.de
Homepage: www.hws-startupfuture.de

ISSN 1612-0884
ISBN 3-910102-76-X

JEL-Klassifikation C80, Z00

Printed in Germany

Inhaltsverzeichnis

1 Einleitung

1.1 Das Projekt „Data Mining Engineering”

Wissen als Unternehmensressource gewinnt zunehmend an Bedeutung. Unternehmen verfügen im Allgemeinen über riesige Datenbanken. Das in diesen Datenbanken „versteckte” Wissen wird aber kaum genutzt. Als eines *der* Wissenschaftsgebiete, die sich mit Wissensextraktion beschäftigen, hat sich das *Data Mining* etabliert.

Die Arbeitsgruppe „KiWi - Künstliche Intelligenz in der Wirtschaftsinformatik” am Fachbereich Wirtschaft der Hochschule Wismar untersucht den effizienten Einsatz von Methoden der Künstlichen Intelligenz - insbesondere auch des Data Mining - zur Lösung wirtschaftswissenschaftlicher Problemstellungen.

Als ein Anwendungsgebiet kristallisierte sich das Financial Engineering heraus. Die Wettbewerbssituation der Kreditinstitute hat sich in den letzten Jahren (nicht nur) in Deutschland zunehmend verschärft. Für den Erfolg eines Kreditinstitutes ist es überlebenswichtig, eine langfristige und profitable Kundenbindung aufzubauen. Dafür sind unter anderem eine individuelle Betreuung sowie passende Produkte (Konditionen, Banking-Software etc.) Voraussetzung. Es wird vermutet, dass sich diese und andere Faktoren im Zahlungsverhalten eines Kunden widerspiegeln. Falls dem so ist, dann lassen sich mit Hilfe von Data-Mining Modelle entwickeln, die eine negative Entwicklung, wie z. B. drohende Kundenabwanderung oder ineffiziente (und für die Bank kostspielige) Nutzung der Banking-Software, rechtzeitig erkennen.

In dem vom Ministerium für Bildung des Landes Mecklenburg-Vorpommern geförderten Projekt „Data Mining Engineering” wird in Zusammenarbeit mit der HypoVereinsbank (ehemals Vereins- und Westbank) und dem MedienHaus Rostock untersucht, welche Methoden sich zur Analyse von Daten aus dem Zahlungsverkehr der Bank eignen. Dabei kann auf eine umfangreiche Datenbank zurückgegriffen werden, die über einen Zeitraum von insgesamt 6 Jahren die monatlich kumulierten Umsätze aller Geschäftskunden umfasst.

Ziel ist, anhand der Umsatzentwicklung der Konten eines Kunden eine Veränderung im Geschäftsumfeld zu erkennen. Auf dieser Grundlage kann ein Kreditinstitut rechtzeitig Maßnahmen zur Sicherung des Kundenbestandes ergreifen. Mögliche Szenarien sind:

- Entwicklung eines Modells zur Erkennung signifikanter Abweichungen im Zahlungsverkehr eines Kunden anhand des bisherigen Zahlungsverlaufes, um z. B. einer drohenden Kundenabwanderung rechtzeitig entgegenzuwirken.
- Untersuchung des Zahlungsverhaltens zur Beurteilung der Nutzung der Banking-Software des Kunden und des Online-Zugangs, um beiderseitig ein effizientes Arbeiten zu fördern. So kann z. B. die Eignung einer Software für einen Kunden beurteilt und ggf. ein passenderes Produkt angeboten werden.
- Das übergeordnete Ziel ist eine bessere Auswertung vorhandener Daten über Kunden, um eine aktivere und individuellere Betreuung zu gewährleisten und somit die Kundenbindung zu erhöhen.
- Das wissenschaftliche Forschungsziel ist, die Eignung klassischer Data-Mining-Verfahren für die Zeitreihenanalyse zu untersuchen und notwendige Schritte der Datenvorverarbeitung zu entwickeln.

1.2 Datengrundlage - Transaktionsdaten

Der bargeldlose Zahlungsverkehr wird durch Banken mit Hilfe von Computer-Systemen durchgeführt. Die bei jeder Transaktion anfallenden Daten werden gespeichert und können zu einem späteren Zeitpunkt wieder abgerufen werden. Der konkrete Datensatz einer Transaktion kann sich dabei von Kreditinstitut zu Kreditinstitut unterscheiden, weist aber immer folgende Informationen auf:

- Das von der Transaktion betroffene Konto;
- Die den Vorgang autorisierende Person, z. B. der Kontoinhaber oder bevollmächtigte Mitarbeiter einer Firma;
- Datum und Uhrzeit der Transaktion;
- Die Höhe des Umsatzes;
- Die Art der Transaktion, Gutschrift oder Lastschrift;
- Die Art des Vorgangs, z. B. Überweisung oder Einzahlung.

Durch Verknüpfung mit einer Kundendatenbank lassen sich alle weiteren Informationen hinzufügen, die das Kreditinstitut über einen Kunden besitzt. Interessante Informationek können z. B. sein:

- Daten des Kontoinhabers (z. B. Name, Firma, Branche, Ort, Land);

- Weitere Konten des Kunden, wenn vorhanden;
- Das Kundensegment oder die Kundengruppe (z. B. Großkunden, Privatkunden);
- Das verwendete Online-Banking Produkt;
- etc.

Werden die einzelnen Transaktionen zeitlich geordnet, lässt sich für jeden Kunden und für jedes Konto ein Transaktionsprofil erstellen. Da die wirtschaftliche Tätigkeit eines Kunden immer Einfluss auf sein Zahlungsverhalten hat, wird vermutet, dass sich für jeden Kunden bzw. jedes Konto typische Transaktionsmuster erkennen lassen, wie z. B. periodisch schwankende Umsätze bei saisonabhängigen Unternehmen. Wenn ein solches Muster erkennbar ist, kann auch eine auftretende Abweichung im Zahlungsverhalten festgestellt werden. Ist dies der Fall, kann das Kreditinstitut die Verhaltensänderung beurteilen und bei Bedarf z. B. durch einen Kundenbetreuer entsprechend reagieren. Aufgrund der Vielzahl von Konten und Transaktionen ist es notwendig, die Analyse der Transaktionsdaten durch Computerprogramme automatisch durchzuführen.

Die uns zur Verfügung stehenden Daten umfassen einen Zeitraum von insgesamt sechs Jahren. Sie beinhalten die monatlich kumulierten Umsätze je Kunde, Konto und Vorgang. Aus den Daten lassen sich ca. 406.000 Zeitreihen über jeweils ein Jahr bilden.

Aus Gründen des Datenschutzes liegen uns die Daten nicht vollständig und nur in anonymisierter Form vor. Die Attribute umfassen die Kundennummer, Kontonummer, Vorgangsart, Kundensegment sowie die monatlich kumulierten Umsätze. Auf die anderen der oben aufgeführten Informationen haben derzeit nur unsere Projektpartner Zugriff. Daher ist eine genauere Auswertung unserer Ergebnisse nur durch manuelle Recherche in dem Informationssystem der Bank vor Ort möglich.

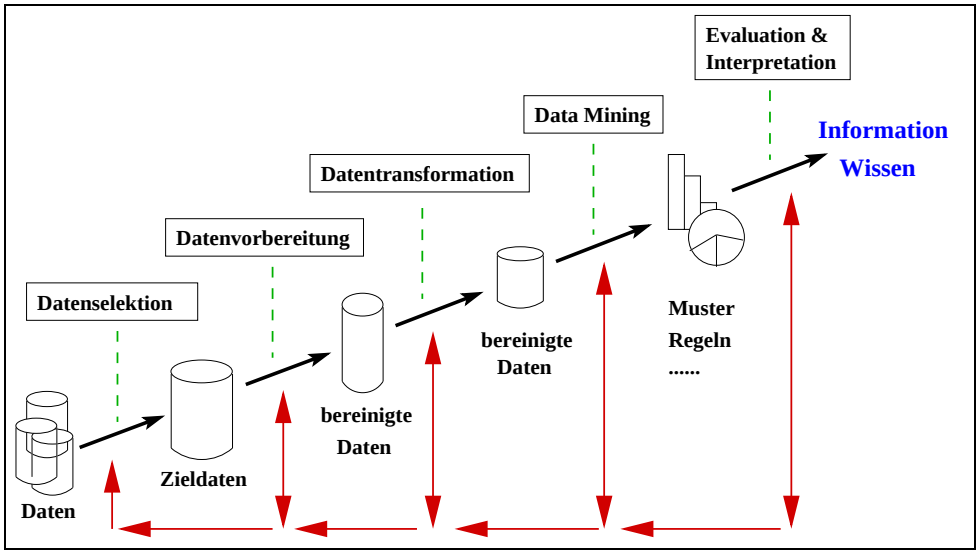
1.3 Data Mining

Als ein Teilgebiet der Wissensextraktion aus Datenbanken (Knowledge Discovery in Databases - KDD) befasst sich *Data Mining* mit der Gewinnung von (neuem) Wissen aus vorhandenen Daten. Die in den Daten implizit enthaltenen Informationen werden durch Anwendung verschiedener Data-Mining-Verfahren extrahiert und in explizites und damit nutzbares Wissen umgewan-

delt. Dieses Wissen kann z. B. durch Regeln oder Diagramme ausgedrückt und zur Lösung von Problemstellungen herangezogen werden.

Als „Data Mining" wird sowohl die Anwendung der Methoden auf Daten, als auch das Forschungsgebiet bezeichnet. Data Mining ist ein stark experimentelles Forschungsgebiet. Dazu gehört sowohl die Erforschung und Entwicklung der Data-Mining-Methoden selbst als auch die Untersuchung, wie diese in der Praxis angewandt werden können. Data Mining als Anwendung ist ein Prozess, der in mehreren Phasen abläuft, wie in Abbildung 1 dargestellt.

Abbildung 1: Data Mining Prozess



Quelle: [FPSSU96].

Eine strikte Trennung einzelner Phasen ist dabei kaum möglich, da sie sich teilweise überschneiden und aufeinander aufbauen.

- Datenvorverarbeitung (Aufbereitung, Kodierung),
- Analyse,
- Datennachbearbeitung (Auswertung, Visualisierung),
- Interpretation.

Vor der eigentlichen Analyse der Daten ist eine *Datenvorverarbeitung* notwendig. Neben der inhaltlichen Bereinigung (Eliminieren von Fehlern, fehlenden Werten oder Ausreißern; Auswählen von Untermengen, Test- und Trainingsdaten) müssen die Daten in eine Form gebracht werden, die durch die

gewählte Data-Mining-Methode verarbeitet werden kann, zum Beispiel durch Normalisierung, Skalierung, Intervallbildung oder Transformation.

Die *Datenerhebung* als Teil der Datenvorverarbeitung hat dafür Sorge zu tragen, dass die einen Sachverhalt beschreibenden Informationen auch in der Datenmenge enthalten sind und andererseits irrelevante, aber signifikante Informationen die Datenmenge nicht verfälschen. Ebenso können ungenaue, fehlende oder falsche Werte das Ergebnis beeinträchtigen oder unbrauchbar machen. Die Datenvorverarbeitung hat entscheidenden Einfluss auf das Ergebnis der eingesetzten Methoden, da durch die Wahl ungeeigneter Vorverarbeitungsverfahren die Daten verfälscht werden können. Es ist notwendig, bei der Kodierung die Daten korrekt und reproduzierbar abzubilden und dabei deren implizite Eigenschaften und Beziehungen zu berücksichtigen.

Die *Datenanalyse* bezieht sich auf die eigentliche Anwendung der Data-Mining-Methoden. Je nach Zielstellung finden Methoden aus den Bereichen Klassifikation, Clustering, Prognose oder Assoziation Anwendung. Um die für die zu analysierenden Daten optimale Methode aus dem jeweiligen Bereich zu ermitteln, sind umfangreiche Experimente mit verschiedenen Methoden und Parametern notwendig.

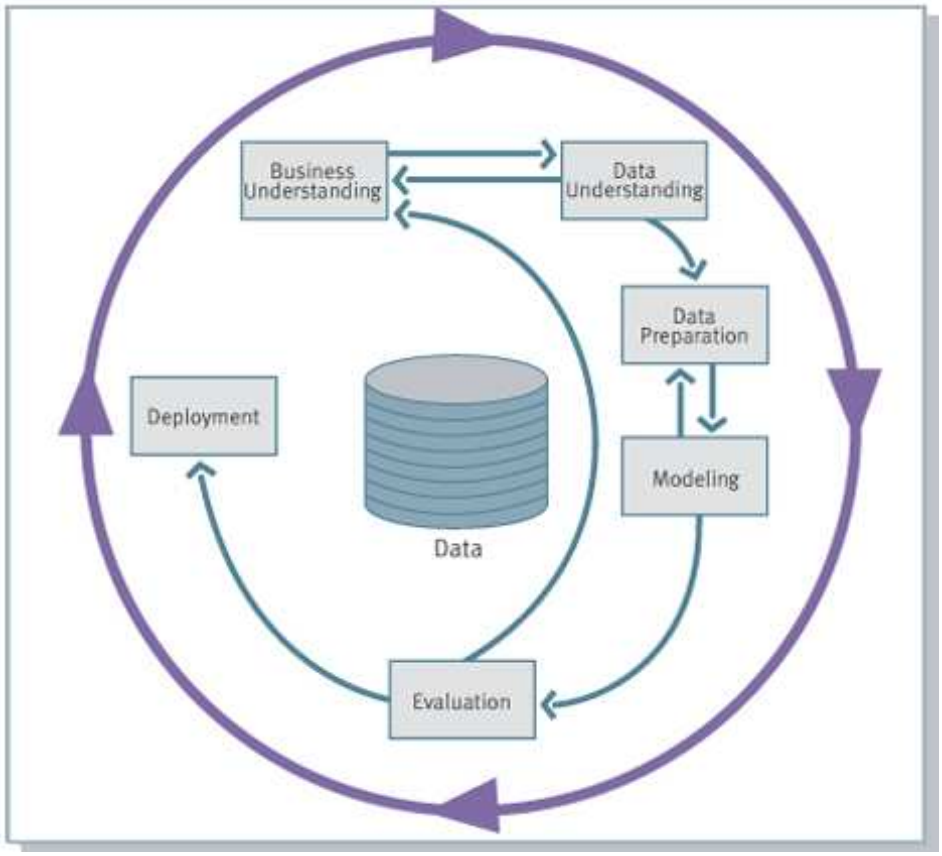
Die *Validierung* dient der Überprüfung des Data-Mining-Ergebnis auf tatsächliche Korrektheit. Dies geschieht meist durch Anwenden des erstellten Modells auf ausgewählte Testdaten.

Datenvorverarbeitung (inkl. Datenerhebung) sowie die Validierung sind dem Data-Mining-Prozess vor- bzw. nachgelagert. Beide sind stark anwendungsbezogen und daher nur bedingt dem Data Mining selbst zuzuordnen. Jedoch haben sie großen Einfluss auf den Erfolg des Data Mining.

Das CRISP-DM[1] Vorgehensmodell bietet einen ähnlichen Ansatz. Es wurde von dem CRISP-DM Konsortium, dem unter anderem Daimler-Chrysler, SPSS und NCR angehören, entwickelt. Das Modell beschreibt Vorgehensweise, Methoden und Ziele beim Data Mining vor allem aus Sicht des Projektmanagements in Unternehmen.

[1] CRISP-DM: CRoss-Industry Standard Process for Data Mining, siehe http://www.crisp-dm.org

Abbildung 2: CRISP-DM Modell



Quelle: http://www.crisp-dm.org

2 Methoden und Algorithmen

Im Folgenden werden Methoden und Algorithmen dargestellt, die bei der Analyse von Transaktionsdaten angewendet werden können. Dabei wird sowohl auf Algorithmen zur Erstellung von Modellen als auch für die Datenaufbereitung eingegangen.

2.1 Datenvorverarbeitung

Vor dem Einsatz der Data-Mining-Algorithmen ist es meist notwendig, die Daten in eine Form zu bringen, die für den jeweiligen Algorithmus geeignet

ist. Die üblichen Verfahren zur Angleichung von Wertebereichen von Datensätzen wie Normierung oder Skalierung sind in der Literatur ausgiebig beschrieben und werden als bekannt vorausgesetzt. Aus diesem Grund wird im Folgenden nur auf ein einziges - für unsere Experimente jedoch maßgebliches - Verfahren zur Signaltransformation, die Fourier-Transformation, eingegangen.

2.1.1 Grundlagen der Fourier-Transformation

Eine Fourier-Reihe[2] besteht aus einer Anzahl von Sinus- und Cosinusschwingungen. Durch additive Überlagerung ist es möglich, jede stetige periodische Funktion annähernd nachzubilden. Die Frequenzen der einzelnen Funktionen sind dabei ganzzahlige Vielfache der Grundfrequenz ω=2π/T, wobei T dem Betrachtungszeitraum entspricht. Die resultierende Schwingung ergibt sich aus:

$$f(t) = \sum_{n=0}^{N} (A_n cos(n\omega t) + B_n sin(n\omega t))$$

Da eine Sinusfunktion einer phasenverschobenen Cosinusfunktion entspricht, lässt sich die Fourier-Reihe auch als Cosinus- und Phasenspektrum darstellen:

$$f(t) = \sum_{n=0}^{N} A_n cos(n\omega t + \varphi_n)$$

Die Fourier-Reihe einer Schwingung oder Funktion kann durch die Fourier-Transformation erzeugt werden. Je nach Eigenschaft der zu zerlegenden Funktionen kommen dabei spezielle Varianten der Fourier-Tranformation zum Einsatz.

2.1.2 Diskrete Fourier-Transformation

Die diskrete Fourier-Transformation (DFT) ist Voraussetzung für viele Anwendungen in der digitalen Signalverarbeitung. Sie erlaubt die Transformation von Signalen, die durch Abtastung als Reihe diskreter reeller Messwerte

[2] Entwickelt von Jean Baptiste Joseph Fourier

vorliegen, vom Zeitspektrum in das Frequenzspektrum. Für die Erkennung eines Signalanteils der Frequenz *n* sind mindestens *2n+1* Abtastpunkte notwendig.

Die Diskrete Fourier-Transformation entspricht der komplexen Multiplikation des Signalvektors mit dem Abtastsignal, in diesem Fall der Sinusfunktion für den Frequenzanteil und der Cosinusfunktion für den Phasenanteil. Der Fourier-transformierte Vektor *F* eines gegebenen Signalvektors *V* der Länge *N* ergibt sich für den Sinusanteil der Frequenzen *f=[0,...,N*-1] aus:

$$F_{real}[f] = \frac{1}{N}\sum_{n=0}^{N}[V_n * sin(\omega f)]$$

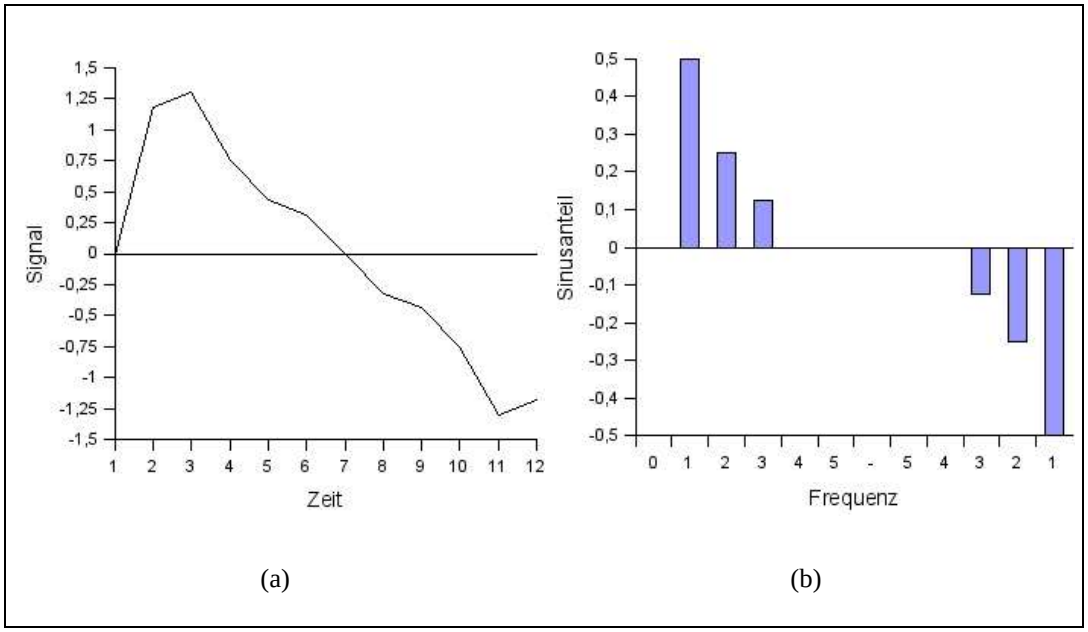
und analog dazu für den Cosinusanteil aus:

$$F_{imag}[f] = \frac{1}{N}\sum_{n=0}^{N}[V_n * cos(\omega f)]$$

Der resultierene Vektor enthält die Sinus- und Cosinusanteile als komplexe Fourier-Koeffizienten. Diese können leicht durch Trigonometrie in Paare [Amplitude, Phase] überführt werden. Dabei ist zu beachten, dass die Koeffizienten invers symmetrisch sind, da der Signalvektor überabgetastet wird. F_0 entspricht dabei einer vertikalen Verschiebung des Signals und ist genau einmal vorhanden, während sich die Sinus- und Cosinusanteile für $n = [1, ..., (N-1)/2]$ aus $F_n + \|F_{N-n}\|$ zusammensetzen.

Beispiel: Die Abbildung 3a zeigt eine Messreihe mit 12 Werten im Zeitspektrum. Die maximal detektierbare Frequenz in diesem Intervall beträgt *(12/2)-1=5* Schwingungen. Die sich daraus ergebende Fourier-Transformierte ist in Abbildung 3b im Frequenzbereich mit den jeweiligen Sinus-Anteilen dargestellt. Die Symmetrie ist sehr gut erkennbar. Die Cosinusanteile wurden nicht dargestellt, da in diesem Beispiel keine Phasenverschiebungen vorhanden sind.

Abbildung 3: Beispiel Fourier-Transformation



Quelle: Eigene Darstellung.

2.1.3 Algorithmus der Fourier-Transformation

Der folgende Pseudocode bietet einen einfachen Algorithmus zur Berechnung der Fourier-Transformation des Vektors *signal* der Länge *N*.

```
vector S;
complex vector E;
for i=0 to N {
 for j=0 to N {
  angle=i*2*pi/N*j;
  E.real[i]=E.real[i]+S[j]*sin(angle)/N;
  E.imag[i]=E.imag[i]-S[j]*cos(angle)/N;
 }
}
```

Dieser Algorithmus entspricht der im Abschnitt 2.1.2 dargestellten formalen Beschreibung. Da der Aufwand für die Berechnung proportional zu N^2 ist, wird in der Praxis eine modifizierte Version, die Fast-Fourier-Transformation

(FFT), eingesetzt. Bei dieser kann durch Ausnutzung der Symmetrie und Wiederverwendung bereits berechneter Werte der Aufwand auf $N \cdot logN$ reduziert werden. Die FFT ist nur auf Signalvektoren der Länge 2^n anwendbar.

2.2 Ausgewählte Data Mining Algorithmen

In der Literatur ist eine Vielzahl von Data-Mining-Algorithmen beschrieben, von denen viele in der Praxis mit Erfolg eingesetzt werden. Die folgende Auswahl ist auf Clustering-Algorithmen beschränkt, die relativ leicht einzusetzen sind und einen hohen Verbreitungsgrad haben. Für weiterführende Recherchen sei auf [IW01, LC04, Läm03, Alp00, Nak98] verwiesen.

2.2.1 Allgemeine Klassifikation der Algorithmen

Die Anwendungsmöglichkeiten des Data Mining sind außerordentlich vielfältig. Es lassen sich drei grundlegende Formen der Anwendung unterscheiden:

1. Klassifikation,
2. Assoziation,
3. Clustering,
4. Vorhersage.

Bei der *Klassifikation* werden überwachte Lernverfahren eingesetzt. Es wird anhand der Attribute von bereits klassifizierten Objekten ein Klassifikator erzeugt, der in der Lage ist, unbekannte Objekte ebenfalls korrekt zu klassifizieren. Als Klassifikatoren können zum Beispiel Entscheidungsbäume, k-Nearest-Neighbour oder Neuronale Netze zum Einsatz kommen. Typische Anwendungen sind z. B. Klassifikation von Kunden oder die Schrifterkennung.

Die *Assoziation* unterscheidet sich von der Klassifikation dadurch, dass nicht nur die Klasse, sondern die Ausprägungen beliebiger Attribute und Attributkombinationen eines Objektes prognostiziert werden können. Als Methoden kommen hier der a-priori-Algorithmus oder wiederum Neuronale Netze in Betracht. Beispielanwendungen findet man in der Warenkorbanalyse oder beim Wiederherstellen eines verrauschten oder fehlerhaften Pixelmusters anhand eines vorher trainierten Beispiels.

Das *Clustering*, ein unüberwachtes Verfahren, wird zur Bildung von Gruppen (Cluster) einander ähnlicher Objekte aus einer Grundmenge eingesetzt.

Dabei kommen verschiedene multivariate Verfahren zum Einsatz. Es kann verwendet werden, um die Attribute herauszufinden, die wesentliche Merkmale einer Gruppe darstellen oder durch die sie sich von anderen Gruppen unterscheiden. Bekannte Clustering-Verfahren sind das K-Means-Verfahren und die von Teuvo Kohonen entwickelten Selbstorganisierenden Karten (SOM[3]). Die mathematisch motivierten Support Vector Machines stellen ebenfalls einen erfolgversprechenden Clustering-Ansatz dar.

Die *Vorhersage* ähnelt der Klassifikation. Sie dient der Bestimmung von Zielgrößen anhand gegebener Attributausprägungen. Im Gegensatz zur Klassifikation liefert die Vorhersage jedoch quantitative Werte.

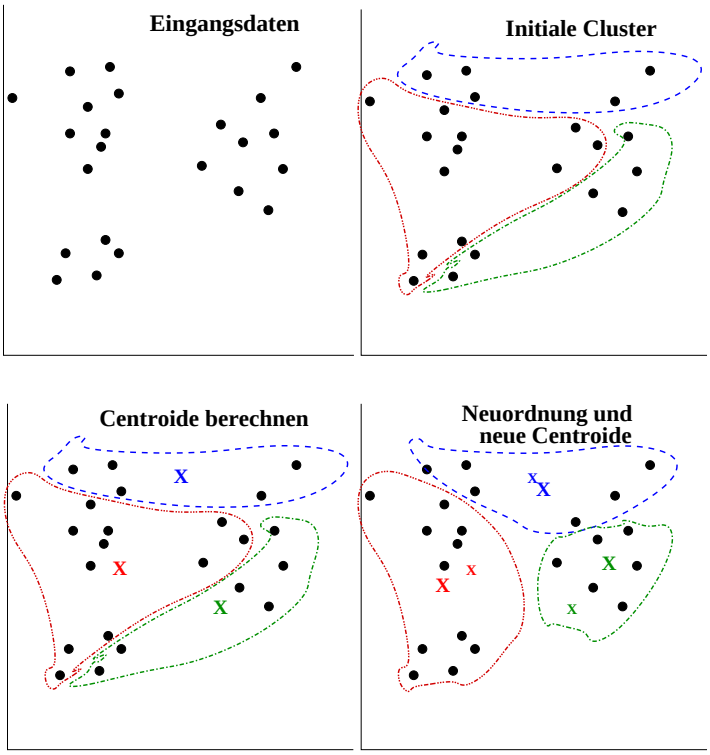
2.2.2 Das Verfahren K-Means

Der K-Means-Algorithmus und dessen Derivat K-Medoid basieren auf der Annahme, dass ähnliche Objekte sich durch einen möglichst geringen Abstand zwischen diesen Objekte auszeichnen. Dazu nutzt man eine gegebene Distanzfunktion $d(a, b)$. K-Means zielt auf die Minimerung des Abstands zwischen dem Objekt und dem Clusterzentrum für Objekte innerhalb eines Clusters. Die Anzahl *k* der zu bildenden Cluster kann beliebig gewählt werden. Voraussetzung für die Anwendung ist, dass die Objekte numerisch als Vektoren der Dimension *n* beschrieben werden können. Der Algorithmus selbst läuft wie folgt ab:

1. Es werden *k* Clusterzentren gebildet und zufällig innerhalb des Datenraumes platziert.
2. Jedes Objekt wird dem Cluster zugeordnet, zu dessen Clusterzentrum der Abstand $d(a, b)$ minimal ist.
3. Für jeden Cluster wird das Clusterzentrum als Schwerpunkt neu berechnet.
4. Falls sich die Zuordnung der Objekte zu den Clustern geändert hat, wird Schritt 2. wiederholt, ansonsten Ende des Algorithmus.

[3] Self-Organizing Map; auch: Kohonen Feature Map.

Abbildung 4: Clusterbildung durch K-Means



Quelle: Eigene Darstellung.

Der Unterschied zwischen K-Means und K-Medoid besteht darin, dass bei K-Medoid immer das nächstgelegene Objekt als Repräsentant des Clusterzentrums dient, während das Clusterzentrum bei K-Means immer ein eigenständiger Vektor - der Schwerpunkt - ist.

Der K-Means Algorithmus führt auch bei großen Datenmengen sehr schnell zu relativ guten Ergebnissen, er hat jedoch auch einige Nachteile:

- Das Ergebnis hängt sehr stark von der Anzahl der gewählten Cluster sowie deren Initialisierung zu Beginn des Algorithmus ab. Es ist möglich, dass sich leere Cluster bilden, denen dann keine Objekte mehr zugeordnet werden können, da sich kein Clustermittelpunkt mehr berechnen lässt.

- Es ist nicht garantiert, dass der Algorithmus in endlicher Zeit konvergiert, da Objekte theoretisch beliebig oft den Cluster wechseln können.

2.2.3 Expectation Maximization

Der EM-Algorithmus (Expectation Maximization) baut auf dem K-Means Algorithmus auf mit dem Unterschied, dass die Zuordnung der Objekte zu den Clustern anhand einer Wahrscheinlichkeitsverteilung realisiert wird. Jedes Objekt *O* gehört somit mit einer Wahrscheinlichkeit *w(O,C)* zum Cluster *C*. Die Wahl der Wahrscheinlichkeitsfunktion hat entscheidenden Einfluss auf das Ergebnis, da jedes Objekt die Bildung aller Cluster beeinflusst. Der Algorithmus ist beendet, sobald die Zuordnung aller Objekte zu den Clustern hinreichend genau ist, d. h. der Abstand der Zugehörigkeitsmaße (likelihood) zwischen den Clustern für jedes Objekt ein zuvor gewähltes Maß übersteigt, oder eine zuvor definierte Anzahl von Iterationen erreicht ist.

2.2.4 Neuronale Netze

Künstliche neuronale Netze sind eine stark idealisierte Nachbildung der Funktionsweise von biologischen Neuronalen Netzen, wie beispielsweise Gehirne oder Nervensysteme. Die Verarbeitung von Informationen erfolgt nicht durch komplexe Algorithmen, sondern vielmehr durch sehr einfache Einheiten, die allerdings in großer Zahl vorhanden sind und untereinander Informationen austauschen.

Analog zu ihren biologischen Vorbildern bestehen künstliche neuronale Netze aus Neuronen und einem Verbindungsnetzwerk. Die Informationsverarbeitung erfolgt mit Hilfe der Propagierungsfunktion, die für jedes Neuron den Aktivierungszustand anhand der von anderen Neuronen eingehenden Signale und der Verbindungsgewichte errechnet. Ein daraus abgeleitetes Ausgabesignal wird dann an andere Neuronen über das Verbindungsnetzwerk weitergeleitet. Es kann als gerichteter Graph mit gewichteten Kanten angesehen werden. Es definiert, welche Neuronen miteinander kommunizieren. Die Gewichte dienen der Hemmung oder Verstärkung von Signalen. Zusammen mit dem Schwellwert (Bias) der Neuronen, der bestimmt, ab welchem Aktivierungsgrad ein Neuron aktiv wird und Signale aussendet, sind es die Gewichte, in denen das „Wissen" des Netzes gespeichert wird. Durch diese verteilte Repräsentation des Wissen sind neuronale Netze relativ unempfindlich gegenüber unvollständigen und verrauschten Eingabemustern.

Die Neuronen sind typischerweise in Schichten angeordnet. Auf die Neuronen der Eingabeschicht werden die zu verarbeitenden Daten in geeigne-

ter Kodierung als Aktivierungswerte übertragen. Diese Aktivierungen werden dann durch das Netz verarbeitet. Das „Ergebnis" kann nach vollständiger Propagierung an den Aktivierungen der Neuronen in der Ausgabeschicht abgelesen werden.

Neuronale Netze sollten immer dann Anwendung finden, wenn andere Lösungsansätze wie z. B. algorithmische Lösungen oder regelbasierte Wissensdarstellungen nicht erfolgreich waren oder nur zu sehr aufwändigen Resultaten geführt haben.

Unüberwachte Lernverfahren finden überall dort Anwendung, wo Datenmengen nach unbekannten Regeln zu klassifizieren sind, was beim Data Mining meist der Fall ist. Ein gutes Beispiel für die Anwendung unüberwachten Lernens sind die von Teuvo Kohonen entwickelten Selbstorganisierenden Karten, auch Feature Maps genannt.

Eine SOM ist ein Neuronales Netz, dessen Neuronen typischerweise als zweidimensionale Gitterstruktur angeordnet sind, wie in Abbildung 5 dargestellt. Die Anzahl der Neuronen beeinflusst die Genauigkeit und die Fähigkeit zur Generalisierung der SOM. Jedes Neuron besitzt einen gewichteten Vektor *W* als Verbindung mit den Neuronen der Eingabeschicht, deren Anzahl *n* der Variablen im Eingaberaum entspricht:

$$W_i = [w_{i,1}, ..., w_{i,n}]$$

Das Trainieren der SOM beginnt mit der Initialisierung der Gewichtsvektoren durch Zufallszahlen im Intervall [-1,1]. Während des Trainings wird ein zufällig gewählter Eingabevektor *I* mit dem Gewichtsvektore *W* jedes Neurons der Kartenschicht verglichen. Das Neuron, dessen Gewichtsvektor der Eingabe am ähnlichsten ist, wird das Gewinnerneuron (BMU, Best Matching Unit). Als Abstandsmaß wird in der Regel die Euklidische Distanz verwendet.

$$d(I, W) = \sqrt{\sum_{i=0}^{n} (I_i - W_i)^2}$$

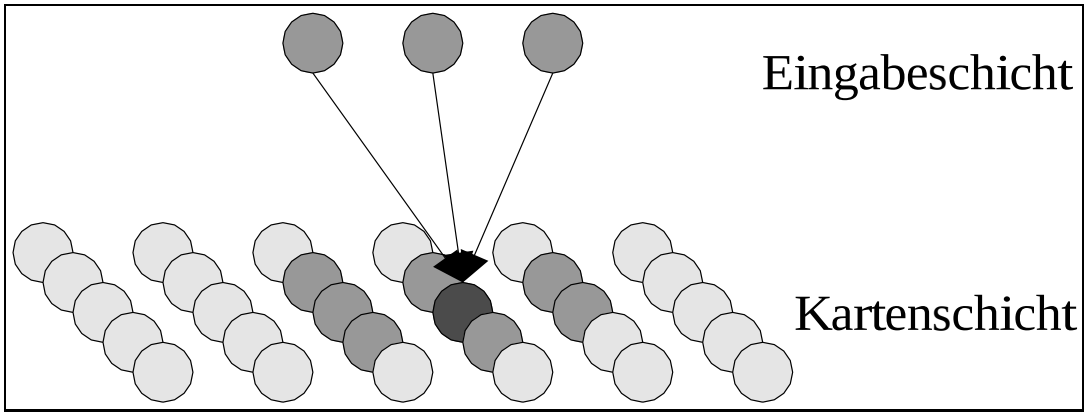
Der Gewichtsvektor *W* des Gewinnerneurons wird anschließend so verändert, dass er dem Eingabevektor *I* ähnlicher wird. Das gleiche gilt für die Neuronen innerhalb eines Radius um das Gewinnerneuron. Der Grad der Beeinflussung wird durch den Lernfaktor η und die Distanzfunktion *h* (meist Gauss-

Funktion) bestimmt, die normalerweise nach jedem Trainingsschritt reduziert werden[Zel00, Läm03]:

$$W_j(t+1) = W_j(t) + \eta(t)h_{cj}(t)[X(t) - W_j(t)]$$

Eine SOM hat die Eigenschaft einer topologieerhaltende Transformation eines hochdimensionalen Eingaberaumes auf eine niedrigere Dimension. Ähnliche Eingabevektoren werden dabei auf benachbarte Neuronen projeziert. Da meist zweidimensionale SOMs zum Einsatz kommen, ist eine Visualisierung der Ergebnisse relativ einfach.

Abbildung 5: Aufbau einer SOM



Quelle: Eigene Darstellung.

3 Experimente im Projekt

Die Experimente wurden auf Basis von ca. 406.000 Zeitreihen, die aus den uns zur Verfügung gestellten Transaktionsdaten errechnet wurden, durchgeführt. Die Experimente wurden zu einem großen Teil von studentischen Hilfskräften durchgeführt.

Die Datensätze umfassen die in Tabelle 1 aufgeführten Felder.

Tabelle 1: Datensatz der Zeitreihen

Feld	Beschreibung	Primärschlüssel
knd	Kundennummer	X
knt	Kontonummer	X
pro	Vorgangs-Nummer	X
ks	Kundensegment	
M1..M12	Monatliche Umsätze	

3.1 Erster Ansatz: Clustering der Originaldaten

3.1.1 Datenvorverarbeitung

Zunächst sind die für das Clustering zu verwendenden Attribute zu identifizieren. Diese sollen die Verhaltenseigenschaften des durch die jeweiligen Datensätze repräsentierten Instanz (Kunde, Konto) widerspiegeln. Alle anderen Attribute würden das Ergebnis verfälschen und sind folglich zu entfernen.

Die Attribute *knd* und *knt* sind nicht für das Clustering geeignet, da die enthaltenen Ausprägungen keinen Bezug zu den jeweiligen Umsätzen haben. Konto- und Kundennummern werden nicht nach festen Kriterien vergeben und sind in den uns zur Verfügung stehenden Daten nur in anonymisierter Form enthalten.

Das Attribut *ks*, das jeden Kunden einem Kundensegment zuordnet, ist ebenfalls wegzulassen. Die Zuordnung erfolgt durch Mitarbeiter der Bank nach festgelegten Kriterien, wie z. B. der Unternehmensgröße, die jedoch keinen direkten Einfluss auf den Verlauf der Umsätze haben.

Das Attribut *pro*, das die jeweilige Art des Vorgangs als Binärschlüssel enthält, hat ebenfalls keinen Einfluss auf den Umsatzverlauf.

Allerdings ist es denkbar, dass ein Zusammenhang zwischen bestimmten Umsatzverläufe und Vorgängen oder Kundensegmenten existiert. Daher sind die Attribute *pro* und *ks* für die spätere Analyse der Cluster wichtig. Denn wenn es Zusammenhänge gibt, ist dies an einer signifikanten Häufung bestimmter Vorgangs-Arten bzw. Kundensegmente in den Clustern zu erkennen.

Für die Clusteranalyse wurden Merkmalsvektoren gebildet, die aus jeweils zwölf Attributen für die monatlich kumulierten Umsätze eines Jahres M1 bis M12 bestehen. Die Zeitreihen wurden Jahresweise gruppiert und auf

Tabelle 2: Übersicht über die Clusterbildung

Daten	Datensätze	Rechenzeit (Min.)	Anzahl Cluster
1997	8410	81	8
1998	16780	318	14
1999	20132	37	2
2000	23404	287	11
2001	98052	281	2
2002	117386	336	2
2003	122477	351	2

den Wertebereich [0,1] normiert.

3.1.2 Clustering

Das Clustering mittels EM-Algorithmus wurde mit WEKA Version 3.4.3 mit folgenden Parametern durchgeführt:

Parameter	Wert
debug	false
maxIterations	100
minStdDev	0.000001
numClusters	-1 (Automatisch)
seed	100

3.1.3 Ergebnisse

Eine Übersicht der entstandenen Cluster ist in den nachfolgenden Tabellen dargestellt. Tabelle 2 zeigt eine Übersicht über die Anzahl der entstandenen Cluster pro Jahr. Eine Häufigkeitsverteilung der Datensätze innerhalb der Cluster ist in den Tabellen 3 und 4 aufgeführt.

Für die Analyse der entstandenen Cluster wurde eine grafische Darstellung aller Cluster durchgeführt. Drei Cluster sind beispielhaft in Abbildung 6 dargestellt. Es ist zu erkennen, dass die Clusterbildung hauptsächlich vom Vorhandensein eines Umsatzmaximums zu einem bestimmten Zeitpunkt beeinflusst wird. Eine derartige Spitze, wie in Abbildung 6a dargestellt, zeigt

Tabelle 3: Häufigkeitsverteilung nach Cluster/Jahr abs.

Cluster	1997	1998	1999	2000	2001	2002	2003
0	495	956	3627	1453	84195	16711	108548
1	438	930	16505	1423	13857	100675	13929
2	1258	834	-	2615	-	-	-
3	597	980	-	5770	-	-	-
4	796	819	-	2763	-	-	-
5	372	1912	-	1327	-	-	-
6	937	3689	-	1421	-	-	-
7	3517	1068	-	1299	-	-	-
8	-	879	-	1392	-	-	-
9	-	1050	-	1659	-	-	-
10	-	376	-	2282	-	-	-
11	-	846	-	-	-	-	-
12	-	882	-	-	-	-	-
13	-	1559	-	-	-	-	-
$\sum$	8410	16780	20132	23404	98052	117386	122477

sich bei Zeitreihen aller Jahre auch in anderen Clustern zu jeweils unterschiedlichen Zeitpunkten. In den Darstellungen anderer Cluster wie 6b und 6c sind ähnliche Maxima zwar weniger ausgeprägt, aber dennoch deutlich zu erkennen. Eine andere signifikante Form der Umsatzkurven lässt sich bei keinem Cluster feststellen.

Eine derartige Clusterbildung ist für unsere Zwecke nicht geeignet. Die deutliche Gruppierung von Datensätzen anhand von Maxima lässt sich auch ohne Data-Mining-Verfahren realisieren. Des weiteren wird die Form der Umsatzkurve offensichtlich nur unzureichend berücksichtigt. So wäre z. B. die Bildung von Clustern mit deutlichem Sinus-förmigem Verlauf zu erwarten, wie er bei saisonabhängien Unternehmen vorkommt.

3.2 Zweiter Ansatz: Fourier-transformierte Daten

Unsere Experimente zeigen, dass die direkte Verwendung der Zeitreihen ungeeignet ist, um Cluster mit in der Form ähnlichen Zeitreihen zu erzeugen. Die

Tabelle 4: Häufigkeitsverteilung nach Cluster/Jahr rel.

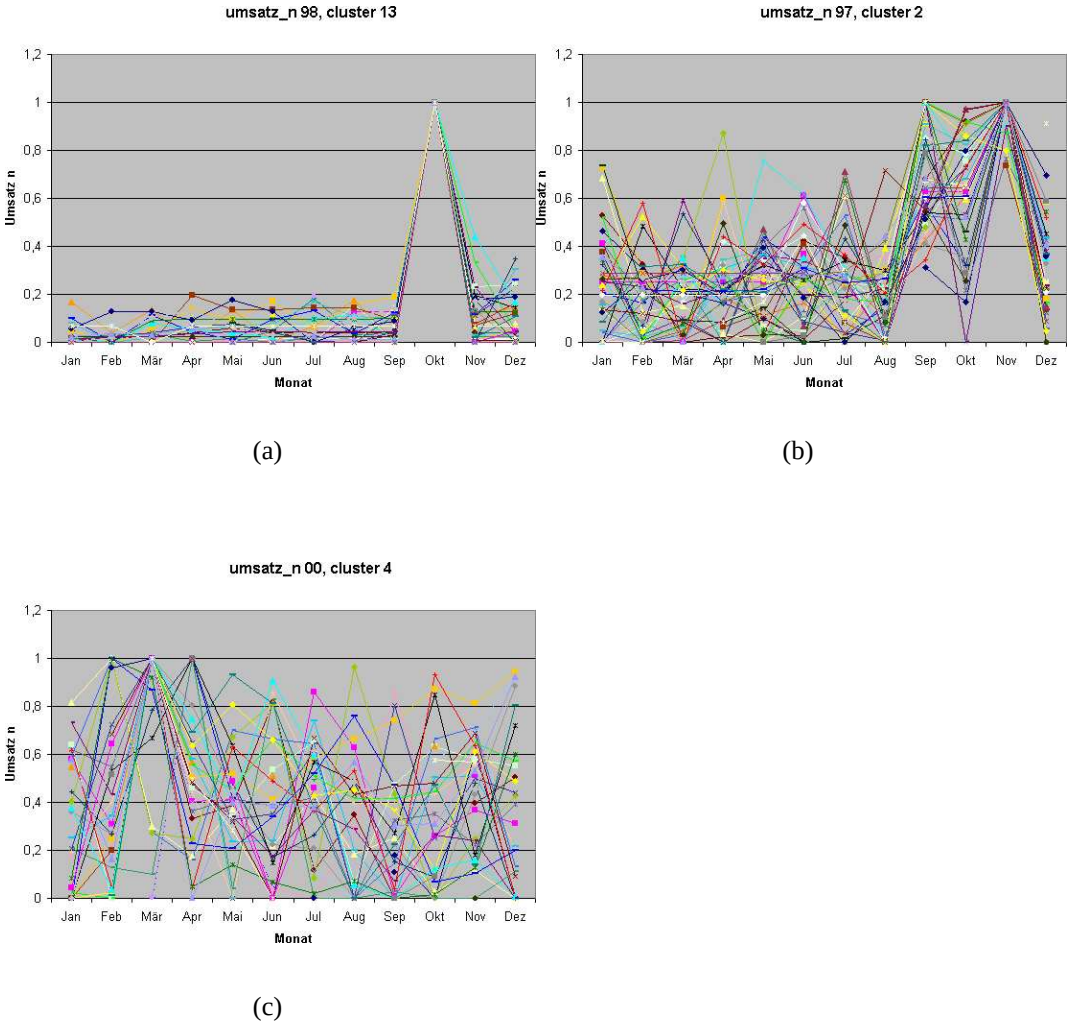
Cluster-Nr.	1997	1998	1999	2000	2001	2002	2003
0	5,89	5,70	18,02	6,21	85,87	14,24	88,63
1	5,21	5,54	81,98	6,08	14,13	85,76	11,37
2	14,96	4,97	-	11,17	-	-	-
3	7,10	5,84	-	24,65	-	-	-
4	9,46	4,88	-	11,81	-	-	-
5	4,42	11,39	-	5,67	-	-	-
6	11,14	21,98	-	6,07	-	-	-
7	41,82	6,36	-	5,55	-	-	-
8	-	5,24	-	5,95	-	-	-
9	-	6,26	-	7,09	-	-	-
10	-	2,24	-	9,75	-	-	-
11	-	5,04	-	-	-	-	-
12	-	5,26	-	-	-	-	-
13	-	9,29	-	-	-	-	-

verwendeten Verfahren benutzen Abstandsmaße, die eine horizontale Verschiebung in den Daten nicht berücksichtigen, da die Attribute unabhängig voneinander betrachtet werden. Um dennoch ein Clustering zu ermöglichen, ist es notwendig, die Daten in eine andere Darstellung zu transformieren, bei der die Reihenfolge der Attribute nicht berücksichtigt werden muß. Dies lässt sich durch die Fourier-Transformation erreichen.

Mit Hilfe der Fourier-Transformation kann eine Zeitreihe von der Form [Zeitpunkt, Amplitude] in die Form [Frequenz, Amplitude, Phase] transformiert werden. Jedes Element des Fourier-Vektors beschreibt ein Attribut der Zeitreihe über das gesamte Intervall, wodurch gleichzeitig die Reihenfolge der Elemente des Fourier-Vektors für das Data Mining unwichtig wird.

Die Abbildung 7 zeigt für verschiedene Konten die Umsatzkurven und daraus abgeleitete Frequenzspektren. In Abbildung 7a ist für beide Konten ein ähnliches Verhalten sowohl an den Umsatzkurven als auch am Frequenzspektrum leicht erkennbar. Beide Konten zeigen relativ konstante Umsätze mit einem ausgeprägten Maximum am Jahresende. Für derartige Zeitreihen liefern abstandsbasierte Clustering-Verfahren gute Ergebnisse. Beide Daten-

Abbildung 6: Visualisierung ausgewählter Cluster



(a) (b)

(c)

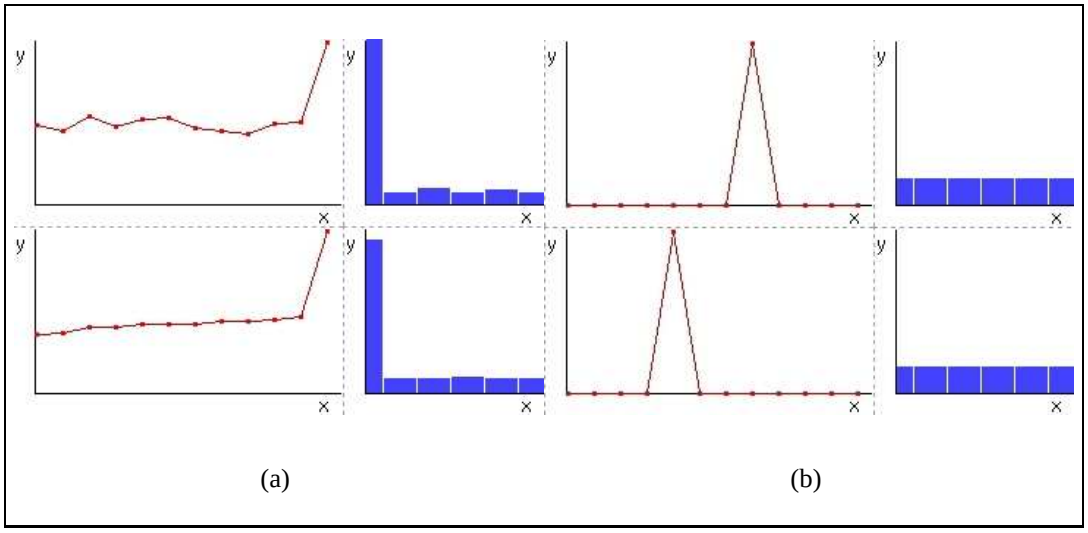
Quelle: Eigene Darstellung.

sätze würden mit hoher Wahrscheinlichkeit ein und demselben Cluster zugeordnet.

Wenn es sich jedoch um ähnliche, jedoch zeitlich Versetzte Umsatzverläufe handelt, so wie bei den in Abbildung 7b dargestellten Konten mit jeweils genau einer Transaktion zu jeweils verschiedenen Zeitpunkten, wird ein abstandsbasiertes Clustering-Verfahren die Datensätze als unterschiedlich voneinander betrachten und mit hoher Wahrscheinlichkeit verschiedenen Clustern zuordnen. Erst die Berechnung des Abstands der Frequenzspektren

zeigt, dass die Zeitreihen in der Form einander ähnlich sind. Die horizontale Verschiebung der Kurven resultiert aus unterschiedlichen, hier nicht dargestellten Phasenanteilen bei in der Amplitude ansonst gleichen Frequenzen.

Abbildung 7: Umsatzverläufe und Frequenzspektren



Quelle: Eigene Darstellung.

3.2.1 Datenvorverarbeitung

Die Datenvorverarbeitung erfolgt durch das im Rahmen eines Projektes entwickelte SXML-Data-Mining-System[Wis03]. Aus den Umsatzdaten des Jahres 2001 wurden 98.051 Zeitreihen gebildet. Die Umsätze der Monate Januar bis Dezember wurden dazu in folgenden Schritten vorverarbeitet:

1. Normierung der Umsätze auf das Intervall [0,1],
2. Fourier-Transformation der normierten Umsätze,
3. Transformation des komplexen Fourier-Vektors $F_{(sin,cos)} \Rightarrow F'_{(amp,phase)}$,
4. Bilden des Merkmalsvektors aus dem Real-Anteil von F'_1 ... F'_5.

Die Amplitude von F'_0 wurde nicht verwendet, da diese nur eine Verschiebung der Kurve in der Vertikalen darstellt.

3.2.2 Clustering

Die durch Clustering erzielbaren Ergebnisse sollen beispielhaft anhand von zwei Experimenten dargestellt werden. Die Experimente wurden unter WEKA mit dem K-Means-Algorithmus sowie mit dem SNNS und einer SOM durchgeführt.

K-Means-Clustering

Die Vorgabe von 6 zu bildenden Clustern führte zu der in Tabelle 5 dargestellten Verteilung. Die Tabelle 8 zeigt die zugehörigen Clusterzentren. Zur besseren Visualisierung sind die Amplituden durch Linien verbunden.

Tabelle 5: Verteilung von 98051 Instanzen auf 6 Cluster

Cluster	Instanzen abs.	Instanzen rel.
0	48513	49%
1	16020	16%
2	11926	12%
3	10177	10%
4	6113	6%
5	5302	5%

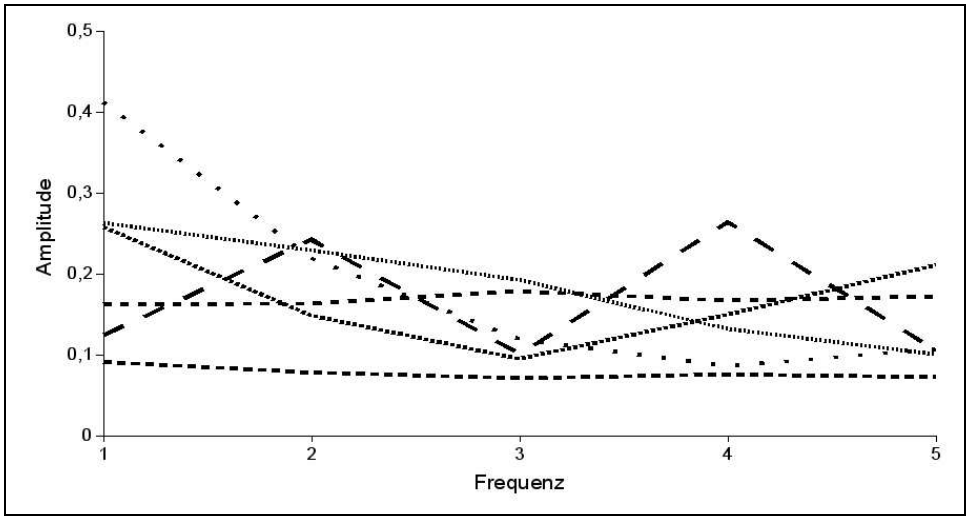
SOM-Clustering

Die Möglichkeiten der Clusterung durch eine SOM soll anhand eines Experimentes gezeigt werden. Dazu wurde eine SOM mit folgenden Parametern erstellt und trainiert:

- Eingabeschicht: 5 Neuronen für die Amplitudenanteile von 1ω bis 5ω,
- Ausgabeschicht: 12 Neuronen, in einem 4x3 Gitter angeordnet,
- 1500 Trainingszyklen,
- Lernfaktor 0.2.

Die Datensätze wurden anhand der Nummer des jeweiligen Gewinnerneurons gruppiert. Das Resultat ist in Abbildung 9 graphisch dargestellt. Die einem Neuron zugeordneten Datensätze sind als Stapeldiagramm abgebildet, um die Eigenschaften der Daten deutlich hervortreten zu lassen. Es ist eine deutliche Gruppierung von Datensätzen mit spezifischen Umsatzverläufen zu erkennen. Besonders auffällig sind hier die konstanten Umsätze aus N0, die deutlich

Abbildung 8: Clusterzentren



Quelle: Eigene Darstellung.

saisonalen Ausprägungen von N4 und N7, sowie die 2 Zahlungsmaxima in N10.

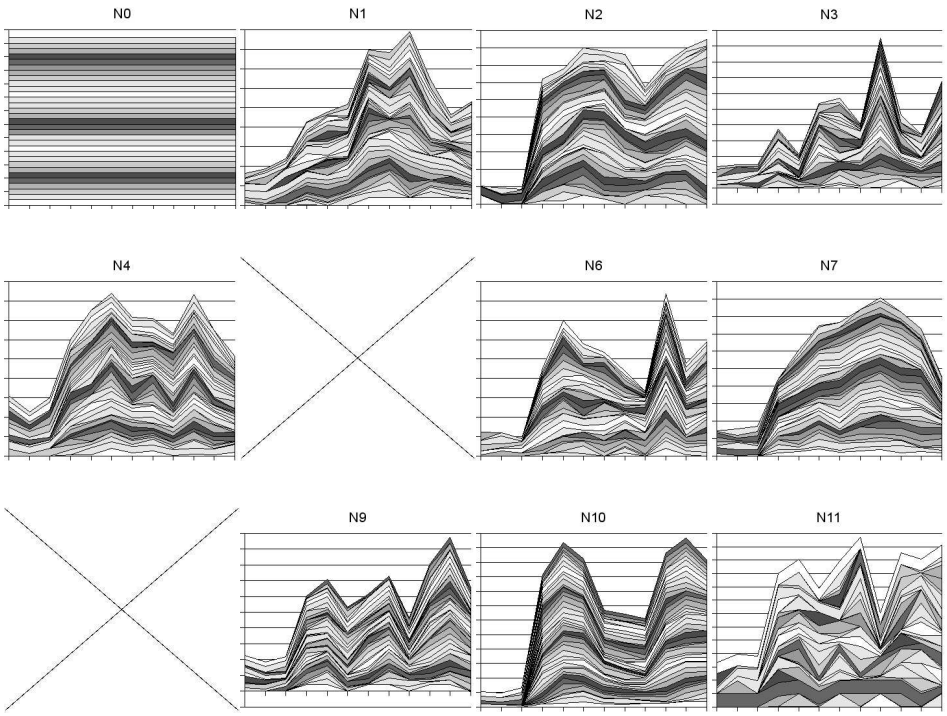
Die meisten Datensätze werden auf die Neuronen N1 (ca. 22.000) und N11 (ca. 51.000) abgebildet. Auf die Neuronen N5 und N8 entfällt kein Datensatz.

3.2.3 Interpretation der Ergebnisse

Es wurden Experimente mit verschiedenen Clustering-Verfahren unter Verwendung der Fourier-transformierten Daten durchgeführt. Obwohl nur die Amplituden für das Clustern verwendet wurden, sind die ersten Ergebnisse sehr vielversprechend.

Es konnten in den entstandenen Clustern repräsentative Umsatzverläufe festgestellt werden, die in Abbildung 9 beispielhaft als Stapeldiagramm dargestellt sind. Durch stichprobenartiges Vergleichen der Umsatzverläufe mit den zugehörigen Kunden und Konten wurde festgestellt, dass teilweise Zusammenhänge mit der Art des jeweiligen Unternehmens bestehen. Die Abbildung 10a zeigt einen Verlauf, wie er häufig bei einem Neukunden nach einer Kontoeröffnung zu finden ist. Bei saisonabhängigen Unternehmen ist eine ausgeprägte periodische Schwankung wie in 10b zu erkennen. Abbildung 10c

Abbildung 9: Datenprojektion auf SOM



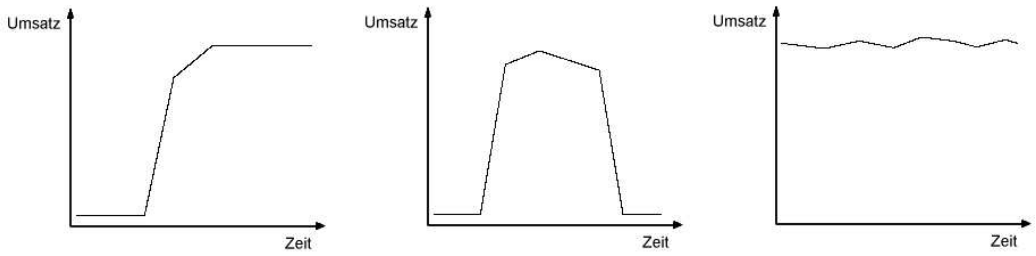
Quelle: Eigene Darstellung.

zeigt eine über das Jahr nur geringen schwankungen unterliegende Umsatzkurve, wie er für Konten mit regelmäßigem Zahlungsein- und Ausgangang auftritt, z. B. bei Immobilienverwaltern oder Gehaltskonten.

3.2.4 Weiterführende Ansätze

Eine genauere Analyse der entstanden Cluster hat ergeben, dass die gezeigte Lösung zwar gute Ergebnisse liefert, aber dennoch Verbesserungen notwendig sind. Es gibt Datensätze, die aufgrund ihres Frequenzspektrums Clustern zugeordnet wurden, obwohl deren Umsatzverläufe sich deutlich von denen des Clusters unterscheiden. In diesen Fällen ist häufig eine stark ausgeprägte Phasenverschiebung in den Frequenzanteilen der Umsatzkurven zu erkennen, die bei der Clusterung bisher berücksichtigt wurde. Zur Lösung des Problems bieten sich folgende Optionen an:

Abbildung 10: Schematische Darstellung typischer Umsatzverläufe



(a) (b) (c)

Quelle: Eigene Darstellung.

1. Die Anzahl der zu bildenden Cluster wird erhöht. Dadurch wird die Abweichung der Datensätze innerhalb eines Clusters reduziert.
2. Das Phasenspektrum wird in das Clustern einbezogen. Durch geeignete Filter sind nicht-signifikante Phasenanteile zu unterdrücken.

Während Punkt 1. leicht umsetzbar ist, erfordert Punkt 2. deutlich mehr konzeptionellen Aufwand. Denn es ist zu beachten, dass es einerseitzs durchaus notwendig sein kann, die Phasenanteile zu ignorieren, um ein verschiebungsinvariantes Clustern zu ermöglichen. Als Beispiel sei der in Abbildung 10a dargestellte Fall einer Signalspitze genannt. Derartige Zeitreihen werden durch das bisherige Vorgehen sehr gut in einem Cluster zusammengefasst. Die Berücksichtigung des Phasenspektrums würde in diesem Fall zu einem erhöhten Abstand zwischen den Datensätzen führen, was das Clustering negativ beeinflussen kann. Andererseits gibt es Fälle, in denen gerade die Phasenanteile starken Einfluss auf die Form des Umsatzverlaufes haben. Eine Möglichkeit zur Lösung des Problems wäre, das Phasenspektrum in Abhängigkeit vom Frequenzspektrum durch den Einsatz von Filtern unterschiedlich stark zu unterdrücken oder zu verstärken.

3.2.5 Möglichkeiten der Nutzung

Die Experimente haben gezeigt, dass ein Clustering von Umsatzverläufen sinnvoll ist. Anhand der entstandenen Cluster kann ein Klassifikator für pro-

totypische Kunden-Klassen erstellt werden, der zur Klassifikation neuer und bestehender Kunden eingesetzt werden kann. Eine Veränderung im Zahlungsverkehr eines Kunden ließe sich feststellen, indem die Zuordnung eines Kunden/Kontos zu einem Cluster bei aufeinanderfolgenden Zeitintervallen untersucht wird. Eine veränderte Cluster-Zuordnung ist dabei ein Hinweis auf eine mögliche signifikante Änderung des Zahlungsverhaltens eines Kunden. Dies kann automatisch erfolgen und im Fall einer veränderten Cluster-Zuordnung ein Signal auslösen, z.B in Form einer Nachricht an einen Kundenbetreuer, das als Hinweis für eine notwendige, intensivere Befassung mit dem jeweiligen Kunden dient.

4 Zusammenfassung und Ausblick

Unsere Experimente zeigen, dass durch die Verwendung der Fourier-Transformation ein verschiebungsinvariantes Clustering von Zeitreihen durch Data-Mining-Algorithmen und Selbstorganisierende Karten möglich ist. Im konkreten Fall bedeutete dies die erfolgreiche Bildung von Clustern, die Datensätze mit ähnlichen Umsatzverläufen beinhalten.

Aus Gründen des Datenschutzes liegen uns die Umsatzdaten nur unvollständig und in anonymisierter Form vor. Die zur Verfügung stehenden Attribute umfassen die Kundennummer, Kontonummer, Vorgangsart, Kundensegment sowie die monatlich kumulierten Umsätze. Auf die anderen der in Abschnitt 1.2 aufgeführen Informationen haben derzeit nur unsere Projektpartner Zugriff. Daher ist eine genauere Auswertung unserer Ergebnisse nur durch manuelle Einzelrecherche für jeden Kunden oder jedes Konto in dem Informationssystem der Bank vor Ort möglich. Dies ist ein extrem zeitaufwändiger Vorgang. Durch stichprobenartige Untersuchung unserer Ergebnisse vor Ort konnte jedoch die prinzipielle Tauglichkeit unserer Vorgehensweise festgestellt werden.

Im bereits begonnenen Folgeprojekt sollen zum einen weitere Tests durchgeführt werden. Zum anderen sollen Konzepte erarbeitet werden, die eine zumindest Semi-Automatisierung der Datenanalysen mittels Data Mining erlauben und somit eine Integration in die Prozessabläufe der Bank gestatten. Dies betrifft z. B. die Anbindung an die bestehende Datenbank der HypoVereinsbank und die Nutzbarkeit der Analyse-Techniken durch Mitarbeiter der Bank.

Literatur

[Alp00] ALPAR, Paul: *Data Mining im praktischen Einsatz*. München : Vieweg, 2000

[CLW05a] CLEVE, Jürgen ; LÄMMEL, Uwe ; WISSUWA, Stefan: *Data Mining auf zeitabhängigen Daten – Kundenanalyse im Bankbereich*. Wismar : Hochschule Wismar, 2005

[CLW05b] CLEVE, Jürgen ; LÄMMEL, Uwe ; WISSUWA, Stefan: Data Mining on Transaction Data. In: DR. NEJDET DELENER, Dr. Chiang-Nan C. (Hrsg.): *Global Markets in Dynamic Environments*. Lissabon : GBATA, 2005

[FPSSU96] FAYYAD, Usama M. ; PIATETSKY-SHAPIRO, Gregory ; SMYTH, Padhraic ; UTHURUSAMY, Ramasamy: *Advances in Knowledge Discovery and Data Mining*. MIT Press, 1996

[IW01] I.H. WITTEN, E. F.: *Data Mining*. München : Hanser, 2001

[LC04] LÄMMEL, Uwe ; CLEVE, Jürgen: *Lehr- und Übungsbuch Künstliche Intelligenz*. Leipzig : Fachbuchverlag Leipzig, 2004

[Läm03] LÄMMEL, Uwe: *Data Mining mittels künstlicher neuronaler Netze, WDP No. 7*. Wismar : Hochschule Wismar, 2003

[Nak98] NAKHAEIZADEH, G.: *Data Mining - Theoretische Aspekte und Anwendungen*. Heidelberg : Physica-Verlag, 1998

[Wis03] WISSUWA, Stefan: *Data Mining und XML - Modularisierung und Automatisierung von Verarbeitungsschritten, WDP No. 12*. Wismar : Hochschule Wismar, 2003

[Zel00] ZELL, Andreas: *Simulation Neuronaler Netze*. München : Oldenbourg, 2000

Online Ressourcen:
WEKA: http://www.cs.waikato.ac.nz/~ml/weka/, last visited 2005-29-03.
CRISP-DM: http://www.crisp-dm.org, last visited 2005-30-06.

WDP - Wismarer Diskussionspapiere / Wismar Discussion Papers

Heft 11/2003	Dietrich Nöthens/Ulrike Mauritz: IT-Sicherheit an der Hochschule Wismar
Heft 12/2003	Stefan Wissuwa: Data Mining und XML. Modularisierung und Automatisierung von Verarbeitungsschritten
Heft 13/2003	Bodo Wiegand-Hoffmeister: Optimierung der Sozialstaatlichkeit durch Grundrechtsschutz – Analyse neuerer Tendenzen der Rechtsprechung des Bundesverfassungsgerichts zu sozialen Implikationen der Grundrechte -
Heft 14/2003	Todor Nenov Todorov: Wirtschaftswachstum und Effektivität der Industrieunternehmen beim Übergang zu einer Marktwirtschaft in Bulgarien
Heft 15/2003	Robert Schediwy: Wien – Wismar – Weltkulturerbe. Grundlagen, Probleme und Perspektiven
Heft 16/2003	Jost W. Kramer: Trends und Tendenzen der Genossenschafts-entwicklung in Deutschland
Heft 01/2004	Uwe Lämmel: Der moderne Frege
Heft 02/2004	Harald Mumm: Die Wirkungsweise von Betriebssystemen am Beispiel der Tastatur-Eingabe
Heft 03/2004	Jost W. Kramer: Der Einsatz strategischer Planung in der Kirche
Heft 04/2004	Uwe Sassenberg: Stand und Möglichkeiten zur Weiterentwicklung des Technologietransfers an der Hochschule Wismar
Heft 05/2004	Thomas Gutteck: Umfrage zur Analyse der Kunden des Tourismuszentrum Mecklenburgische Ostseeküste GmbH
Heft 06/2004:	Anette Wilhelm: Probleme und Möglichkeiten zur Bestimmung der Promotioneffizienz bei konsumentengerichteten Promotions
Heft 07/2004:	Jana Otte: Personalistische Aktiengesellschaft
Heft 08/2004	Andreas Strelow: VR-Control – Einführung eines verbundeinheitlichen Gesamtbanksteuerungskonzepts in einer kleinen Kreditgenossenschaft
Heft 09/2004	Jost W. Kramer: Zur Eignung von Forschungsberichten als einem Instrument für die Messung der Forschungsaktivität
Heft 10/2004	Jost W. Kramer: Geförderte Produktivgenossenschaften als Weg aus der Arbeitslosigkeit? Das Beispiel Berlin
Heft 11/2004	Harald Mumm: Unterbrechungsgesteuerte Informationsverarbeitung
Heft 12/2004	Jost W. Kramer: Besonderheiten beim Rating von Krankenhäusern
Heft 01/2005	Michael Laske/Herbert Neunteufel: Vertrauen eine „Conditio sine qua non“ für Kooperationen?
Heft 02/2005	Nicole Uhde: Rechtspraktische Probleme bei der Zwangseinzie-

hung von GmbH-Geschäftsanteilen – Ein Beitrag zur Gestaltung von GmbH-Satzungen

Heft 03/2005 Kathrin Kinder: Konzipierung und Einführung der Prozesskostenrechnung als eines Bestandteils des Qualitätsmanagements in der öffentlichen Verwaltung

Heft 04/2005: Ralf Bernitt: Vergabeverfahren bei öffentlich (mit)finanzierten sozialen Dienstleistungen

Heft 05/2005: Jost W. Kramer: Zur Forschungsaktivität von Professoren an Fachhochschulen am Beispiel der Hochschule Wismar

Heft 06/2005 Harald Mumm: Der vollständige Aufbau eines einfachen Fahrradcomputers

Heft 07/2005: Melanie Pippig: Risikomanagement im Krankenhaus

Heft 08/2005: Yohanan Stryjan: The practice of social entrepreneurship: Theory and the Swedish experience

Heft 09/2005: Sebastian Müller/Gerhard Müller: Sicherheits-orientiertes Portfoliomanagement

Heft 10/2005: Jost W. Kramer: Internes Rating spezieller Kundensegmente bei den Banken in Mecklenburg-Vorpommern, unter besonderer Berücksichtigung von Nonprofit-Organisationen

Heft 11/2005: Rolf Steding: Das Treuhandrecht und das Ende der Privatisierung in Ostdeutschland – Ein Rückblick –

Heft 12/2005: Jost W. Kramer: Zur Prognose der Studierendenzahlen in Mecklenburg-Vorpommern bis 2020

Heft 13/2005: Katrin Pampel: Anforderungen an ein betriebswirtschaftliches Risikomanagement unter Berücksichtigung nationaler und internationaler Prüfungsstandards

Heft 14/2005: Rolf Steding: Konstruktionsprinzipien des Gesellschaftsrechts und seiner (Unternehmens-)Formen

Heft 15/2005: Jost W. Kramer: Unternehmensnachfolge als Ratingkriterium

Heft 16/2005: Christian Mahnke: Nachfolge durch Unternehmenskauf – Werkzeuge für die Bewertung und Finanzierung von KMU im Rahmen einer externen Nachfolge –

Heft 17/2005 Harald Mumm: Softwarearchitektur eines Fahrrad-Computer-Simulators

Heft 18/2005: Momoh Juanah: The Role of Micro-financing in Rural Poverty Reduction in Developing Countries

Heft 19/2005: Uwe Lämmel, Jürgen Cleve, René Greve: Ein Wissensnetz für die Hochschule – Das Projekt ToMaHS

Heft 20/2005: Annett Reimer: Die Bedeutung der Kulturtheorie von Geert Hofstede für das internationale Management

Heft 21/2005: Stefan Wissuwa, Jürgen Cleve, Uwe Lämmel: Analyse zeitabhängiger Daten durch Data-Mining-Verfahren