Heuristische Verfahren zur Aggregation addierbarer Zeitreihen

Stefan Pohl
Student der
Hochschule Mittweida
Fachbereich Informatik
mail@s-pohl.de

Dr. S. Steinberg
VKG Verlagvertriebs KG
Sachsenfeld 3-5
20086 Hamburg
s.steinberg@bauerverlag.de

Zusammenfassung

Um die Wirkung von Vertriebstests oder Marketingaktionen in sehr heterogenen Märkten zu analysieren, ist man darauf angewiesen, vergleichbare Test- und Kontrollgruppen bilden zu können.
Das in SAS realisierte Verfahren bietet eine heuristische Lösung für die oben beschriebene Aufgabe. Ein anderer Ansatz besteht in der Anwendung von Clusterverfahren aus dem SAS/Enterprise Miner.
Damit die Ergebnisse unterschiedlicher Clusterverfahren gegenübergestellt werden können, wurde ein neuer SAS/Enterprise Miner-Knoten („Cluster Comparison“) entwickelt.

Keywords: Aggregation, Zeitreihe, Vertrieb, Testszenario, Supervised, Cluster, Vorhersage, Prognose, Optimierung.

1 Einleitung

Ungeachtet der Mehrdimensionalität vieler Probleme und Prozesse besitzen doch die meisten eine starke Abhängigkeit von der Zeit. Daher beschäftigen sich viele wissenschaftliche Bereiche mit der Analyse und Vorhersage von Zeitreihen, um durch Planung und Steuerung besser auf Prozesse eingehen zu können.
Gleichzeitig besteht ein Trend zu oligopolen Märkten, die durch starken Konkurrenzkampf geprägt sind. Allzuoft wird deshalb der meiste Profit in Nischenmärkten gemacht. Diese kennzeichnen sich durch niedrige Umsätze, die hohe Umsatzschwankungen nach sich ziehen. Das erschwert die durch Empirie gestützten Analysen der Vorgänge, die zu Verbesserungen führen sollen.
Um dennoch gesicherte bzw. genauere Aussagen machen zu können, bedarf es der Minimierung von nicht erklärbaren Schwankungen. Dies kann durch Auslöschung dieser Effekte erreicht werden, wenn man entsprechende Zeitreihen zusammenfaßt.
In diesem Beitrag soll ein Verfahren aufgezeigt werden, das durch eine Heuristik die Kombinatorik dieses Problems dermaßen beschränkt, daß eine praxisgerechte Berechnung möglich wird. Zuvor werden Versuche vorgestellt, die eine Abbildung auf bereits vorhandene Prozeduren darstellen.

2 Heinrich Bauer Verlag

Die Bauer Verlagsgruppe ist Europas führender Zeitschriftenverlag. Weltweit erstellen, produzieren und vertreiben ca. 6500 Mitarbeiter 116 auflagenstarke Titel. Die Spanne des Angebots reicht vom stark besetzten Fernsehzeitschriftenmarkt („TV Movie“, „Auf einen Blick“, „TV Hören und Sehen“, ...) über Frauenzeitschriften („Bella“, „Tina“, „Vida“, ...), Jugendzeitschriften („Bravo“-Familie) bis hin zu Special-Interest-Magazinen („Geldidee“, „Wohnidee“, „Praline“, ...). Die Verlagsgruppe ist in Deutschland Marktführer in den Segmenten Programmzeitschriften (Marktanteil 55%, Auflage: 9,8 Mio. Exemplare), Jugendzeitschriften (Marktanteil 39%, Auflage: 1,4 Mio. Exemplare) und unterhaltende Frauenzeitschriften (Marktanteil 35%, Auflage: 4,0 Mio. Exemplare). Daraus resultiert, daß fast jeder zweite Deutsche (50,4%) regelmäßig eine Zeitschrift aus dem Hause Bauer kauft.

3 Vertriebsaktivitäten

Bei der Verbreitung von Zeitschriften entstehen während der gesamten Wertschöpfungskette Daten, die Aufschluß über Zusammenhänge geben und in nachgelagerten Schritten beachtet werden müssen, um die Prozesse zu optimieren. Der letzte Schritt ist hierbei der Vertrieb, wo Informationen gebündelt werden, um eine optimale Distribution der Zeitschriften zu ermöglichen. Gleichzeitig ist der Vertrieb die Stelle an der das unmittelbare (also nicht durch Marktforschung oder Stichproben generierte) Feedback des Kunden merkbar wird. Sämtliche Verkäufe und auch remittierende Ware treffen beim Verlag ein. Systembedingt hat man es hier also bereits mit einem hohen Datenaufkommen zu tun.

3.1 Typisches Problem des Vertriebs

Der Vertrieb hat die Aufgabe Zeitschriften optimal im Markt zu verteilen, also mit möglichst wenig Aufwand möglichst viele Kunden zu erreichen. Dazu hat er verschiedene steuernd wirkende Möglichkeiten, die von Werbung und Einzelhandelpräsenz bis zur Erhöhung oder Verringerung der Auflage reichen.
Die Frage, die sich hierbei immer wieder stellt ist: „Wie können die Auswirkungen von vertrieblichen Aktivitäten nachgewiesen werden?“
Ein Lösungsansatz besteht in der testweisen Änderung von Vertriebsparametern. Das bedingt eine Einteilung der Vertriebsgebiete in Test- und Referenzgebiete, die ein möglichst gleiches und damit vorhersagbares Verhalten haben. Sollten die Testgebiete von diesem Verhalten abweichen, ist die Auswirkung nachgewiesen.
Der Vorhersagefehler muß also möglichst klein sein, da Testauswirkungen immer erst aus dieser Bandbreite heraustreten müssen.

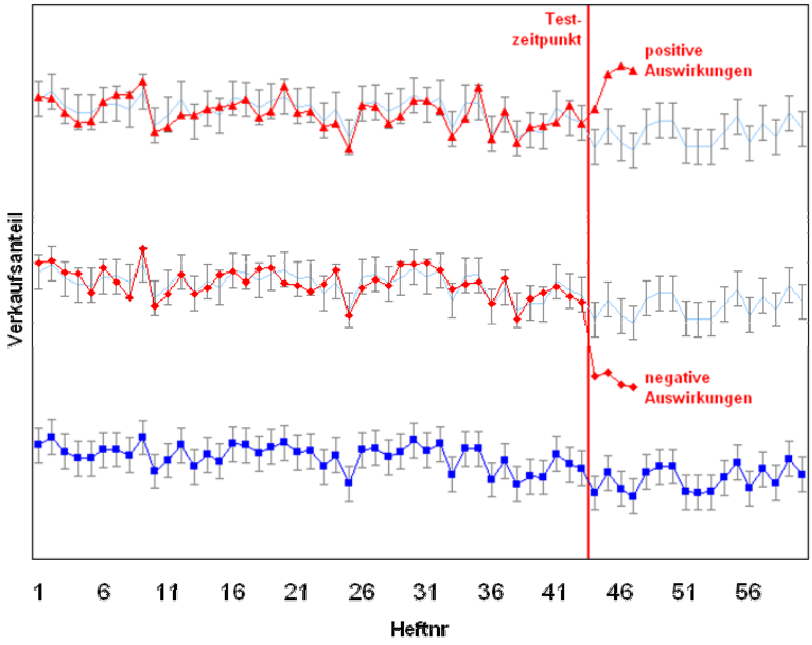

Abbildung 1: Test- und Referenzgebiete in einem Testszenario

Die Illustration eines idealen Testszenarios stellt die Abbildung 1 dar. Hier wird die Referenzzeitreihe in blau mit einer Schwankungsbandbreite über die Testzeitreihen gelegt und über den Testzeitpunkt fortgeschrieben.
Einflüsse, die zu Fehlern in der Vorhersage führen, sind vielfältig und reichen von den konkreten Inhalten jeder Zeitschrift, über die Werbewirksamkeit des Titelbildes bis hin zur Regionalität des Einzelhandels. So können Titelthemen zeitlich begrenzte, regionale Verkaufsschlager sein. Die Eigenschaften der Kunden sind unbekannt und sind beliebig geographisch verteilt. Zusätzlich verzerren bisherige Tests die Historie und sorgen für Ausreißer in der Zeitreihe.

3.2 Lösungsansatz

An die Lösung dieser Probleme kann man sich herantasten, indem man Vertriebsgebiete zusammenfaßt, die so ein besser vorhersehbares Verhalten erreichen.
Das Problem läßt sich formal wie folgt definieren: In einer n x m Matrix wird der Mittelwert und die Standardabweichung in jeder Zeile ermittelt. Durch löschen oder addieren von Zeilen wird nun versucht eine neue kleinere Matrix zu erzeugen, in der die Summe eines Schwankungsmaßes wie die Standardabweichung über alle Zeilen hinweg einem Minimum entgegenstrebt. An diese kleinere Matrix können nun Anforderungen hinsichtlich der minimalen Anzahl der Zeilen oder der Aggregationen gestellt werden.

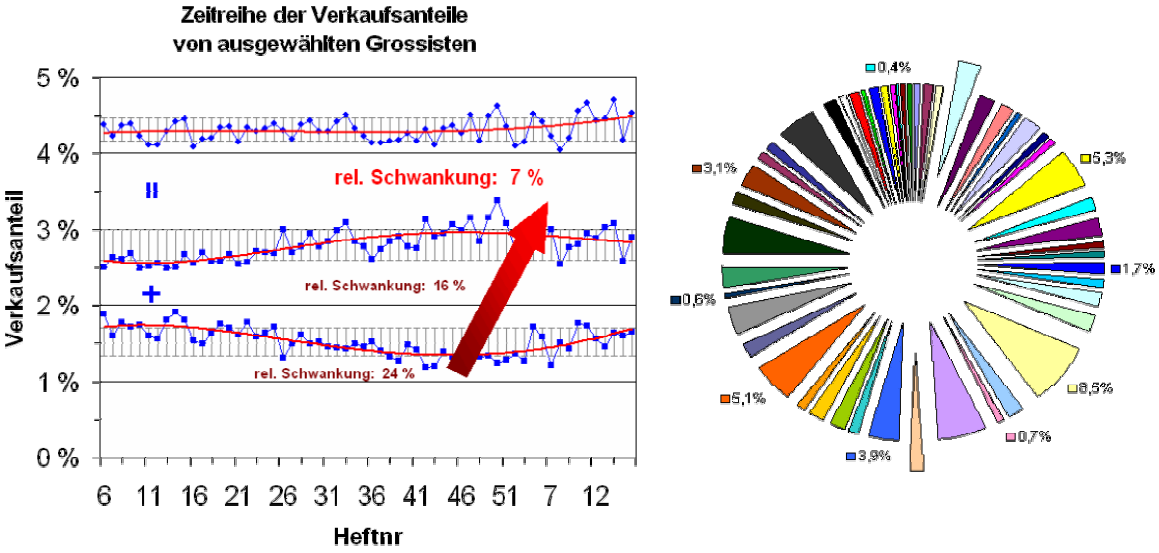

Abbildung 2: Aggregation zweier Zeitreihen

Aus den Verkaufsanteilen der Vertriebsgebiete aus Abbildung 2 wird deutlich, daß es sich hier im Beispiel um ein Wirtschaftsmagazin handelt: Es gibt viele große Anteile (Großstädte) und viele kleine Anteile (ländliche Gegenden).
Nach dem „Gesetz der großen Zahlen“ muß man Vertriebsgebiete nun wie in Abbildung 2 auswählen, so daß sich zeitliche Schwankungen bestmöglichst ausgleichen.

3.3 Abbildung auf Verfahren des SAS/Enterprise Miners

Der Versuch diese Problematik auf Verfahren des SAS/Enterprise Miners abzubilden, durchlief mehrere Stufen.
Eine Zufallsauswahl läßt keine Optimalität erwarten, da die Auswahl nicht zielgerichtet ist. Dennoch können hieraus erzeugte Aggregationen als Benchmark dienen, um andere Verfahren einordnen zu können.
Die Clustering-Verfahren weisen ebenso die Eigenschaft auf, daß sie ungerichtet sind. Sie stellen Algorithmen zur Verfügung, die eine Statik der Elementeigenschaften zugrundelegen. In diesem Problem besitzen zusammengefaßte Vertriebsgebiete aber gänzlich andere Eigenschaften als jedes einzelne.
Das Ziel von klassischen Cluster-Algorithmen besteht in der Herauskristallisierung von Zusammengehörigkeitsbeziehungen zwischen Elementen, die durch die Auswertung von Abstandsmaßen entstehen. Hier werden also in bezug auf dieses Abstandsmaß ähnliche Elemente aggregiert. Diese Vorgehensweise kann zur Lösung des Problems nur beitragen, wenn man für die zu aggregierenden Zeitreihen unterstellt, daß sich gerade die Zufallsschwankungen von im Trend stark ähnlichen Zeitreihen ausgleichen.
Das im folgenden vorgestellte Verfahren lehnt sich vielmehr an die Vorgehensweise bei der Variablenselektion an, die Elemente (Variablen) zielgerichtet in zwei Gruppen („erklärende“ und „redundante“ Variablen) einteilt.

4 Supervised Clustering

Der hier vorgestellte Algorithmus zur Lösung des Problems kann wegen seiner Zielorientierung als „Supervised Clustering“ bezeichnet werden. Er beschränkt sich auf eine vergleichsweise kleine Untermenge der Möglichkeiten Vertriebsgebiete zu aggregieren. Das entspricht der menschlichen Vorgehensweise, immer nur paarweise Vertriebsgebiete zusammenzufassen und dann erneut zu versuchen, diese zu verbessern.

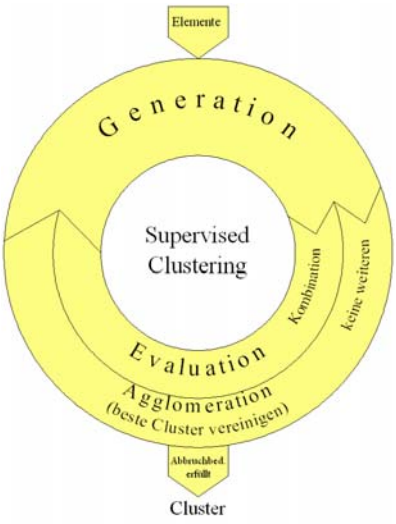

Abbildung 3: Algorithmus (schematisch)

In Abbildung 3 ist der Vorgang angedeutet: Beliebige zu clusternde Einzelelemente werden in Tabellenform (SAS konform) dem Algorithmus zugeführt. Es folgt die Generierung einer Menge von Kombinationsmöglichkeiten, die dann einer Auswertung zugeführt werden. Sind alle in diesem Stadium generierten Möglichkeiten erschöpft, erfolgt die Verschmelzung der Elemente mit dem besten Auswertungsergebnis („scoring“).
Der Vorgang läuft solange fort, bis eine Abbruchbedingung erfüllt ist. Diese kann prinzipiell an jede Aktion geknüpft werden. Im Ergebnis stehen dann Aggregationen der Einzelelemente zur Verfügung.

4.1 Komplexität

Interessant ist eine Betrachtung der Komplexität des Problems. Will man nur 5 Gruppen aus 50 Elementen bilden, benötigt man über 10^{34} Auswertungen, selbst die Einschränkung pro Gruppe genau 10 Elemente zu erhalten, beschränkt die Komplexität nur marginal ($>10^{31}$). Erst das hier vorgestellte Verfahren hat ein Zeitverhalten ($<10^{5}$), das eine solche Untersuchung praktikabel macht.
Selbst aus 120 Elementen 10 Gruppen zu bilden, stellt kein Problem dar, obwohl die Anzahl der Möglichkeiten dann 10^{120} beträgt[1].

[1] Das ist die Zahl, mit der von Experten die Anzahl der möglichen Spielzustände beim Schach nach oben hin abgegrenzt wird.

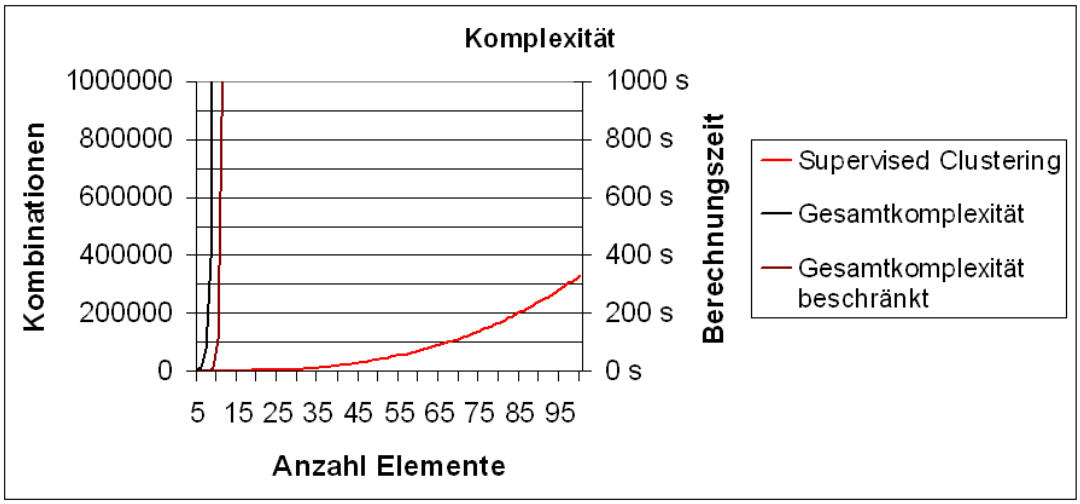

Abbildung 4: Zeitverhalten in Abhängigkeit von der Elementanzahl
(aufgenommen mit Pentium 800 MHz, nichtoptimierter Interpretercode)

Abbildung 4 visualisiert die Zusammenhänge, die im folgenden hergeleitet werden sollen:

Binomialkoeffizient: $b(n,k) := \frac{n!}{(n-k)!\cdot k!}$

Anzahl Entitäten: n

Anzahl Gruppen: g

e..Anzahl Entitäten pro Gruppe bei Gleichverteilung:

$$e(n,g) := \operatorname{ceil}\left(\frac{n}{g}\right) \qquad e(50,5) = 10$$

paarweise Forward-Selektion: $forw(n) := \sum_{i=2}^{n} i\cdot(i-1)$

$forw(50) = 41650$ (siehe Abb. 4)

komplette Backward-Selektion in jeder Gruppe:

$$backw(n,g) := n\cdot g\cdot \sum_{i=1}^{e(n,g)} b(e(n,g),i)$$

$backw(50,5) = 255750$

Komplexität Stepwise: $O(n,g) := forw(n) + backw(n,g)$

$O(50,5) = 297400$

4.2 Implementierung

Aufgrund der unklaren Anforderungen und Nebenbedingungen sowie der Unkenntnis der konkreten Verbesserung erfolgte die Implementierung eines ersten Prototyps in Prolog. Die „logische Programmierung“ war hier das Mittel der Wahl, da sie sich einerseits aus einer Datenbasis, in der beliebige Daten assoziativ speicherbar sind, und andererseits aus Regeln zusammensetzt, die auf diesen Daten arbeiten. So ist eine einfache nachträgliche Definition von einschränkenden Bedingungen möglich.
Bei jeder Aktion des Algorithmus können Entscheidungen unter Einhaltung eines komplexen Regelwerks getroffen werden. Beispiele hierfür sind die Fragen: „Welche Elemente kommen anfänglich in Betracht“, „ist die Zusammenführung zweier Elemente erlaubt“, „wird dabei eine Größenrestriktion für Cluster überschritten“ oder „wann soll das Verfahren abbrechen“.
In SAS stehen ebenfalls die Daten in Form von Datasets im Mittelpunkt, auf denen Prozeduren und Data-Steps arbeiten. Aufgrund dieser Verwandtschaft war eine einfache Portierung nach SAS/Base möglich.
Durch die Implementierung einer stabilen Version in SAS wurde eine nahtlose Integration des Programms in andere Untersuchungen sowie den SAS/Enterprise Miner möglich. Da SAS durch die Nutzung von Datasets in externen Speichern arbeitet und darauf optimierte Operationen anbietet, ist eine hohe Skalierbarkeit gewährleistet.

4.3 Ergebnis

Das Ergebnis der Bemühungen war eine deutliche Fehlerreduktion der Vorhersagen in allen Gruppen.

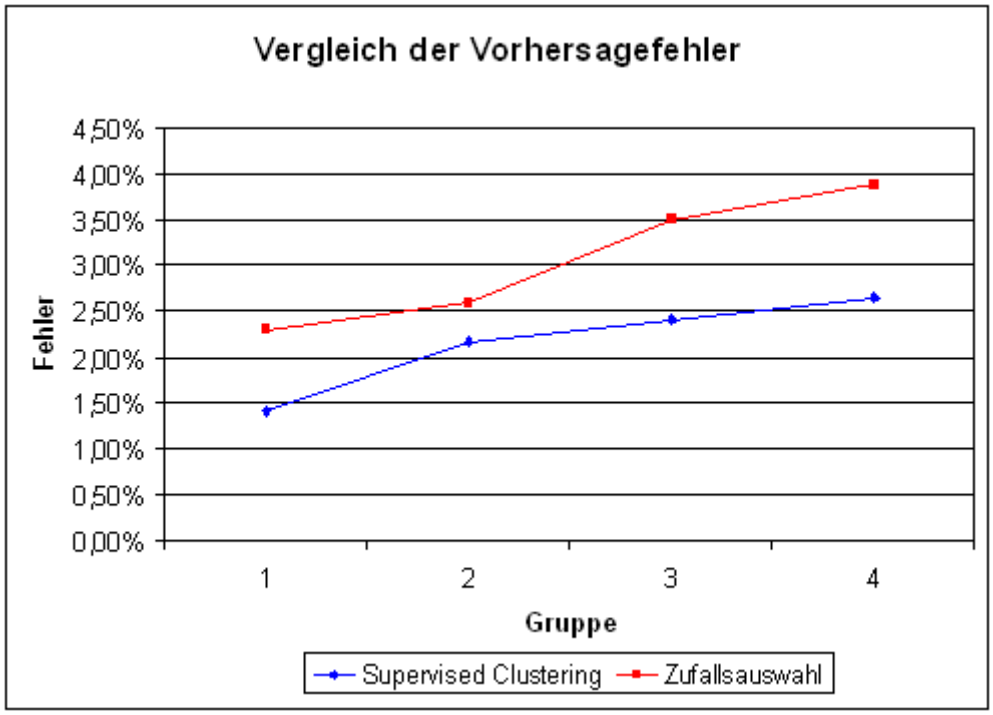

Abbildung 5: Vergleich der Vorhersagefehler

Der Ablauf des Verfahrens gab Auskunft darüber, daß es wenige sich stark ausgleichende Vertriebsgebiete gibt. Das ist in der starken Volatilität und Regionalität der Schwankungen begründet.

5 Vergleich von Klassifikationen

Durch verschiedene Parametrisierungen oder Verfahren entstehen schnell eine Vielzahl von Cluster-Zugehörigkeiten bzw. Klassifikationen. Die Frage stellt sich, inwieweit sie sich voneinander unterscheiden? Oder: „Auf welche Cluster einer anderen Gruppierung verteilen sich die Elemente eines Clusters?“
Zur Darstellung der Verteilungsmöglichkeiten eignen sich Kreuztabellen. Der Anschaulichkeit halber werden hier nur zweidimensionale Kreuztabellen benutzt.

Absolute Anzahl	Cluster ID SOM			
	1	2	3	All
	N	N	N	N
Cluster ID KMEANS				
1	2	1	.	3
2	1	2	1	4
3	.	1	2	3
All	3	4	3	10

Abbildung 6: Kreuztabelle zweier Clusterings

Ein selbstgeschriebener SAS/Enterprise Miner-Knoten erzeugt paarweise Vergleiche aller mit ihm verbundenen Cluster- und Klassifizierungsknoten.
Eine starke Übereinstimmung resultiert in einer Konzentration der Elemente entlang der Hauptdiagonalen. In Abbildung 6 wird in den Schnittpunkten die Anzahl der Elemente angezeigt, wobei aber prinzipiell jede beliebige Aggregatfunktion möglich ist. So ist es oftmals auch interessant die Volumina (Summierung von Elementeigenschaften wie Verkaufsanteile) hinter den Elementanhäufungen zu sehen.

6 Zusammenfassung

Zusammenfassend kann gesagt werden, daß durch das hier vorgestellte „Supervised Clustering“ die breite Testbarkeit bei einem auflagenschwachen, volatilen Titel erst möglich geworden ist. Desweiteren gibt das Agglomerationsverhalten während des Cluster-Vorgangs Auskunft über Eigenschaften der zugrundegelegten Daten und Prozesse.

Weitere Verbesserungen der Vorhersagbarkeit sind durch den Einsatz evolutionärer Algorithmen [4] zu erwarten, da hier hochdimensionale, kombinatorische Räume durchsucht werden können, ohne daß durch Expertenwissen Teile dieses Raumes von vornherein ausgeschlossen werden müssen. Dadurch wäre eine höhere Sicherheit gegeben, ein globales Optimum zu finden.
Der „Cluster Comparison“-Knoten steht durch die vollständige Integration in die Tools des SAS/Enterprise Miners für Wiederverwendbarkeit und schnelle Auswertung von Scorings. Der Überlappungsgrad gibt Hinweise auf die Stabilität der Aggregationen.

Literatur

1. Hastie, T., Tibshirani, R., Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York.
2. Müller H-D. (2000). MBR less & more. In dnv Sonderheft 50 Jahre Presse-Grosso, S. 116-155. Hamburg.
3. Ulf Nilsson, Jan Maluszynski (2000). Logic, Programming and Prolog (2Ed). http://www.ida.liu.se/~ulfni/lpp.
4. Biethahn, Jörg et al. (1998). Betriebswirtschaftliche Anwendungen des Soft Computing. Vieweg, Braunschweig.