Clusteranalyse als Tool zur Selektion und Verarbeitung elektromagnetischer Daten

Stefan L. Helwig
Institut für Geophysik und Meteorologie der Universität zu Köln

1 Einleitung

Bei der Verarbeitung großer Datenmengen ist es oftmals nicht möglich, alle aufgezeichneten Daten im einzelnen anzusehen. Um einen Überblick über die gesamte Datenmenge zu erhalten, ist man auf statistische Parameter wie Mittelwert, Standardabweichung oder Korrelation angewiesen. Diese Werte geben lediglich dann ein korrektes Bild, wenn die zugrundeliegenden Rohdaten normalverteilt sind. Ist das nicht der Fall, oder sind die Daten durch systematische Fehler verzerrt, können Mittelwert und Standardabweichung ein sehr ungenaues Abbild der Gesamtdatenmenge liefern.
Eine Möglichkeit, einen besseren Überblick zu erhalten, kann die Clusteranalyse sein. Unter diesem Begriff werden verschiedene Verfahren zur Sortierung von Daten in Gruppen zusammengefaßt. Diese Verfahren wurden erfolgreich in anderen Wissenschaften wie in der Biologie oder der Psychologie eingesetzt, lassen sich aber auch gut auf Zeitreihen anwenden.

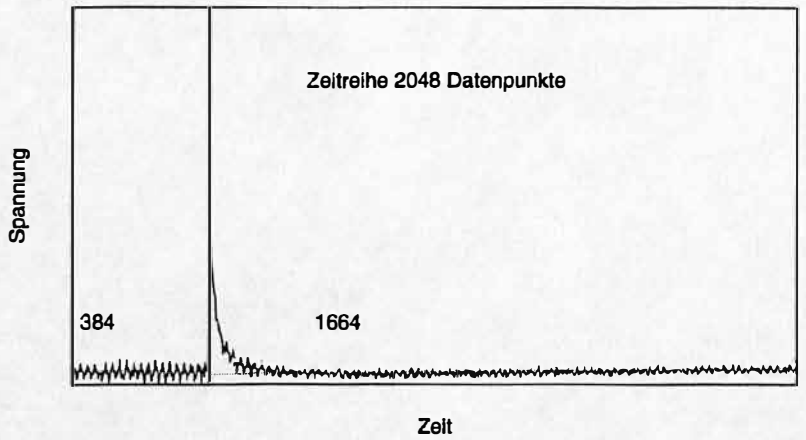

Abbildung 1: *Struktur der mit LOTEM aufgezeichneten Zeitreihen. Alle im weiteren gezeigten Datensätze bestehen aus 384 Datenpunkten vor dem Umschalten und aus 1664 Punkten danach.*

Die im Folgenden betrachteten Zeitreihen sind alle mit der LOTEM-Methode (Strack, 1992) aufgezeichnet worden. Sie bestehen in diesem speziellen Fall jeweils aus 2048 Datenpunkten, von denen 384 vor dem Schaltzeitpunkt des Senders liegen und somit das Nullniveau bilden. Die restlichen 1664 Datenpunkte enthalten das eigentliche Nutzsignal (Abbildung 1).
Bei den in den letzten Jahren durchgeführten LOTEM-Meßkampagnen ist die aufgezeichnete Datenmenge durch Nutzung einer Mehrkanalapparatur enorm gestiegen. An einem Meßtag können ca. 12 Senderpositionen mit jeweils 32 Empfängerpositionen und 100 Einzeltransienten aufgezeichnet werden. Die dabei entstehende Datenmenge von 38400 Transienten bildete die Motivation, nach Möglichkeiten für einen schnellen Überblick über große Datenmengen zu suchen. Um die Daten schon während einer laufenden Kampagne bearbeiten zu können, müssen die benutzten Verfahren sowohl schnell als auch möglichst stark automatisierbar sein.

2 Theorie zweier Clusteralgorithmen

Ziel der Clusteranalyse soll es sein, die vorhandene Datenmenge so aufzubereiten, daß ungewöhnliche Ereignisse aussortiert werden und sofort auffallen. Dazu werden alle mit einer Sender-Empfängerkombination aufgezeichneten Transienten miteinander verglichen und nach Ähnlichkeiten sortiert. Von großer Bedeutung ist hierbei, daß vor der Durchführung der Clusteranalyse periodisches Rauschen möglichst sorgfältig von den Daten entfernt wird, da sonst die Gefahr besteht, daß die Zeitreihen nach der Phase des Rauschens und nicht nach etwaigen Unterschieden in den Nutzsignalen sortiert werden.
Die Theorie zu zwei verschiedenen Clusteralgorithmen wird hier nur kurz angerissen. Eine ausführlichere Darstellung der Algorithmen sowie eine umfangreiche Liste von Referenzen zur Clusteranalyse findet sich bei Steinhausen und Langer (1977).

2.1 Hierarchisches Clustern

Bei diesem Verfahren bildet zunächst jeder Transient ein eigenes Cluster. In einem iterativen Prozeß werden dann jeweils die beiden ähnlichsten Cluster zu einem zusammengefaßt, bis nur noch eine vorgegebene Anzahl von Clustern übrigbleibt.
Die Ähnlichkeit zwischen zwei Clustern (Transienten) wird über die Distanzmatrix bestimmt. Sie wird berechnet durch :

$$d(t_i t_j) = \sum_{k=1}^{m} |t_{ik} - t_{jk}| \quad , \tag{1}$$

wobei m die Anzahl der Datenpunkte im Transienten ist. Die verwendete Metrik muß dem jeweiligen Problem angepaßt sein. Die in Gleichung 1 genutzte Distanz eignet sich gut, um sehr kleine Unterschiede in den Daten stark zu betonen. Sie ist ein Spezialfall der Minkowski-r-Metriken,

$$d(t_i t_j) = \left(\sum_{k=1}^{m} |t_{ik} - t_{jk}|^r \right)^{1/r} \quad , \tag{2}$$

bei denen kleine Differenzen in den Daten mit zunehmendem r eine immer geringere Bedeutung haben.
Unabhängig von der Wahl der Metrik befinden sich auf der Hauptdiagonalen der Distanzmatrix immer Nullen, da jeder Transient zu sich selbst die größte Ähnlichkeit hat. Darüber hinaus ist sie immer symmetrisch, da die Ähnlichkeit des Transienten i zu j genauso groß ist wie die von j zu i.
Im weiteren muß daher nur eine untere bzw. obere Dreiecksmatrix betrachtet werden. Das kleinste Element dieser Dreiecksmatrix gibt die beiden Cluster mit der größten Ähnlichkeit an. Die Inhalte dieser beiden Cluster werden in einem der beiden kombiniert, und Zeile und Spalte des übriggebliebenen Clusters werden gestrichen.
Bis zu diesem Punkt ist der Prozeß des hierarchischen Clusterns mathematisch exakt. Das Problem liegt nun darin, wie die Distanzen des kombinierten Clusters zu den anderen Clustern berechnet werden. Natürlich ist es möglich, die Distanzmatrix vollständig neu zu berechnen. Insbesondere bei großen Datensätzen würde das aber zu unvertretbar langen Rechenzeiten führen. Effektiver ist es, die Distanzen des kombinierten Clusters zu allen übrigen Clustern aus den Distanzen der beiden alten Cluster zu den übrigen zu berechnen. Im einfachsten Fall definiert man die Distanzen des neuen Clusters als den Minimalwert aus den Distanzen der beiden alten und erhält:

$$d_{ij}^{neu} = min\left(d_{ij}^{alt} + d_{ik}^{alt}\right);\ d_{jk}^{alt}\ kleinstes\ Element;\ j < k \tag{3}$$

Nach Lance und Williams (1966) neigt der Clusteralgorithmus bei derartig neubestimmten Distanzen dazu, entferntere Cluster bei der Kombination zu bevorzugen.
Für die LOTEM-Daten hat es sich als besser erwiesen, statt des Minimalwerts der beiden alten Distanzen ihren Mittelwert zu verwenden:

$$d_{ij}^{neu} = \frac{1}{2}\left(d_{ij}^{alt} + d_{ik}^{alt}\right);\ d_{jk}^{alt}\ kleinstes\ Element;\ j < k \tag{4}$$

Sind die neuen Distanzen berechnet, wird wieder das kleinste Element der Distanzmatrix gesucht, und die beiden dazugehörigen Cluster werden kombiniert. Danach werden erneut die Distanzen berechnet und so fort. Dieser Vorgang wird so lange wiederholt, bis nur noch die gewünschte Zahl von Clustern übrigbleibt.

2.2 Clustern Sift and Shift

Eine völlig andere Strategie benutzt das Clusterverfahren Sift and Shift. Die Daten werden beliebig in n Cluster vorsortiert, und die Mittelwerte der Cluster $M(c) : c = 1..n$ werden berechnet.
Durch die Art des Vorsortierens erhalten die Daten bereits eine bestimmte Struktur. Eine Sortierung der ersten x Transienten in das erste Cluster, der nächsten x in das zweite und so weiter würde z.B. eine Änderung der Daten mit der Zeit stärker betonen. In Fällen, in denen eine neutrale Vorsortierung gewünscht ist, können die Transienten nach dem Zufallsprinzip auf die Cluster verteilt werden.
Nach dem Vorsortieren werden alle Transienten gesichtet (sift). Das bedeutet, daß für jeden Transienten i die Abstände a zu den verschiedenen Clustermittelwerten berechnet werden:

$$a_{ic} = \sum_{k=1}^{m} |t_{ik} - M_{ck}| \quad . \tag{5}$$

Im nächsten Schritt wird jeder Transient in das Cluster geschoben (shift), zu dem er die geringste Entfernung hat. Ist das für alle Transienten geschehen, werden erneut die Clustermittelwerte berechnet, und der Prozeß beginnt von vorne. Er wird beendet, wenn in einer Iteration kein Transient mehr das Cluster wechselt.
Beide Verfahren wurden mit verschiedenen LOTEM-Datensätzen getestet und führen zu etwa gleich guten Ergebnissen. Der Vorteil des Hierarchischen Verfahrens ist seine höhere Geschwindigkeit. Bei systematischen Problemen, die bei jedem zweiten Transienten auftreten, wie das bei einigen Messungen beobachtet wurde, hat der Sift and Shift Algorithmus die höhere Genauigkeit.

3 Anwendung auf Felddaten

Die linke Seite der Abbildung 2 zeigt die Mittelwerte dreier Cluster eines E-Feldes, das während der letzten Meßkampagne am KTB aufgezeichnet wurde (Sylvester, 1996). Bei dieser Kampagne nutzten verschiedene Arbeitsgruppen das gleiche Sendesignal. Für ein Experiment zum nichtlinearen IP-Effekt (Bigalke, 1996) wurde, nachdem der Sender für etwa 90 Umschaltvorgänge auf einer konstanten Stromstärke gehalten worden war, das Sendesignal in mehreren Stufen abgesenkt, und es wurden jeweils noch einige Transienten aufgezeichnet.
Der gesamte Datensatz an dieser Station umfaßt etwas mehr als 100 Transienten. Sie wurden nach Entfernen des periodischen Rauschens in vier Cluster gruppiert. Ein Cluster enthält nur einen einzigen Transienten. Es ist daher auf Abbildung 2 nicht dargestellt. Die Unterschiede in den aufgezeichneten Spannungen der anderen sind deutlich zu erkennen.
Die genaue zeitliche Abfolge des Experiments ist im rechten Teil der Abbildung 2 zu erkennen. Aufgetragen ist die Nummer des Clusters, in die ein Transient eingruppiert wurde, gegen die

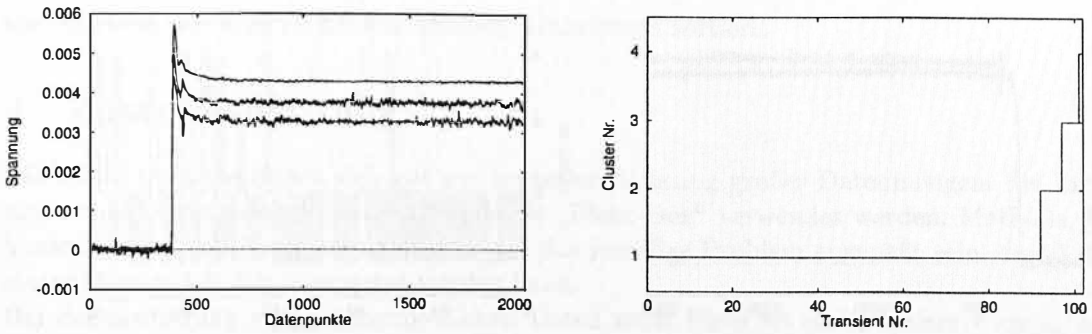

Abbildung 2: *Mittelwerte für drei Cluster und Zugehörigkeit der Einzeltransienten zu den Clustern für eine Station am KTB.*

Nummer des Transienten, die der zeitlichen Reihenfolge beim Aufzeichnen entspricht.
Es ist klar erkennbar, daß die Experimente zum nichtlinearen IP-Effekt nach etwa 90 Umschaltvorgängen des Senders begonnen haben.

3.1 Problemerkennung durch Clusteranalyse

Der auf Abbildung 3 dargestellte Datensatz eignet sich gut, um den Vorteil der Clusteranalyse bei systematischen Problemen zu erläutern. Dargestellt ist der Mittelwert von über 100 Einzeltransienten einer Station. Die mit aufgetragenen Fehlerbalken für die Standardabweichung des Mittelwertes sind so klein, daß sie in dieser Darstellung gar nicht sichtbar werden. Auf den ersten Blick scheint der Datensatz daher eine sehr hohe Wiederholungsgenauigkeit zu haben.

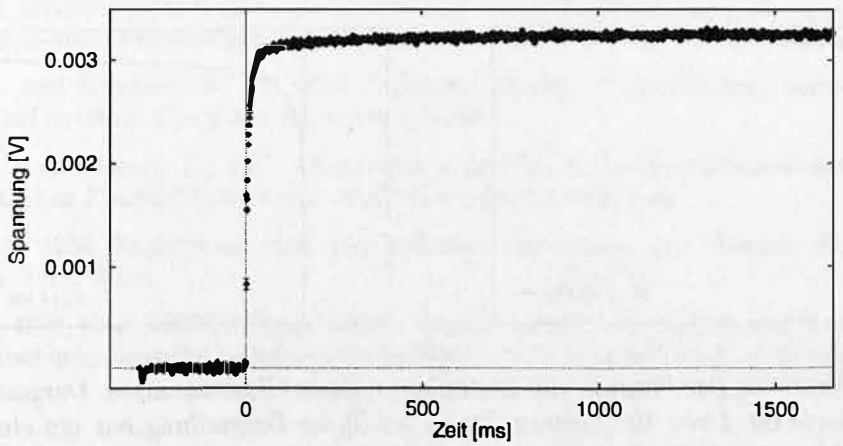

Abbildung 3: *Mittelwert und Standardabweichung von 102 Transienten eines $\dot{H}_z$-Feldes.*

Bei einer Einteilung in Cluster ist aber erkennbar, daß die Transienten mit geraden Nummern in andere Cluster einsortiert werden als die mit ungeraden Nummern (Abbildung 4). Die Clustermittelwerte zeigen deutlich, daß die Hälfte der Daten höhere Spannungswerte annimmt und eine leichte Drift aufweist. Selbst bei Betrachtung von Einzeltransienten sind diese Unterschiede nur sehr schwer festzustellen.

3.2 Vergleich verschiedener Stationen

Eine weitere Möglichkeit, die die Clusteranalyse bietet, ist der Vergleich von Zeitreihen, die an verschiedenen Stationen aufgezeichnet wurden. Hierbei soll die Form der Transienten im Vordergrund stehen. Als Grundlage für die Analyse werden die Mittelwerte aller an einer Station

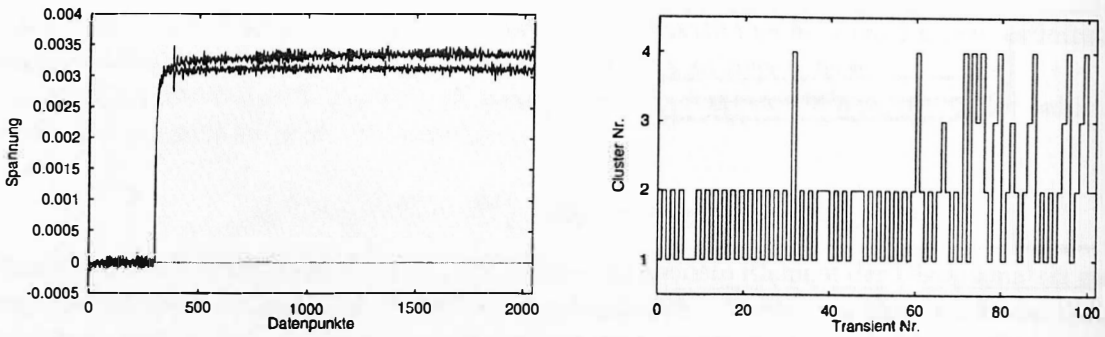

Abbildung 4: *Mittelwerte für vier Cluster und Zugehörigkeit der Einzeltransienten zu den Clustern*

aufgezeichneten Transienten verwendet. Damit die jeweilige Amplitude keinen Einfluß auf das Sortierergebnis hat, müssen die Daten zunächst normiert werden.

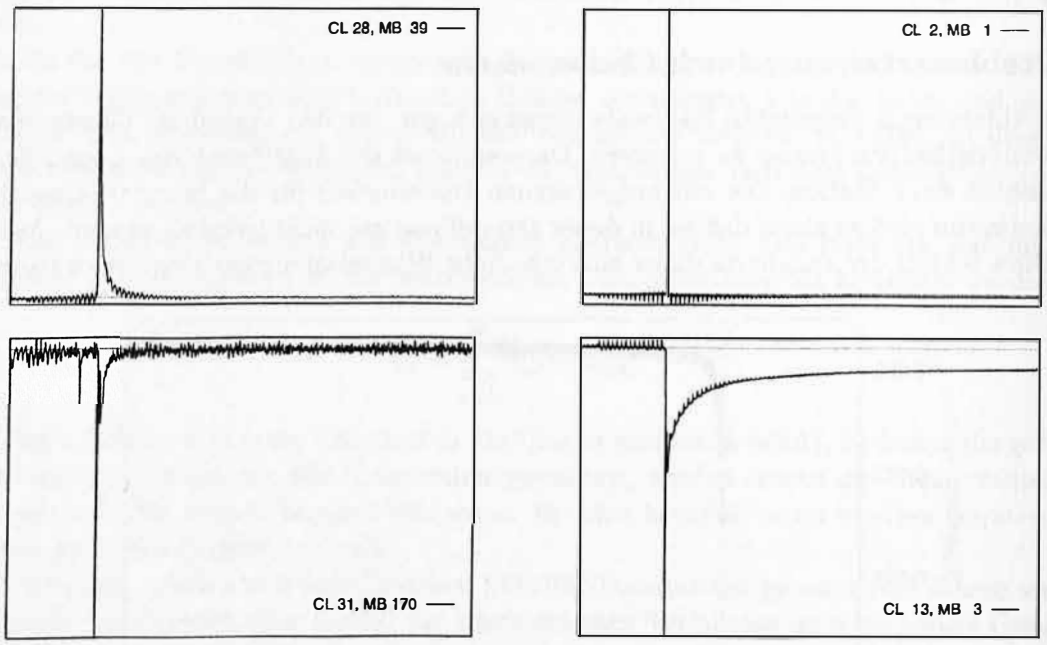

Abbildung 5: *Vergleich der Formen von Zeitreihen mittels Clusteranalyse. Dargestellt sind die Clustermittelwerte für 4 von 40 Clustern. Da es bei dieser Darstellung nur um einen Vergleich der Formen geht, sind bewußt keine Skalen angegeben.*

Abbildung 5 zeigt exemplarisch vier Clustermittelwerte aus einem Vergleich von 400 $\dot{H}_z$-Feldern einer Meßkampagne. Nach dem Normieren wurden die Daten in 40 Cluster gruppiert.
Damit Transienten mit einmaliger Form nicht zu anderen Transienten sortiert werden, darf die Anzahl der vorgegebenen Cluster nicht zu klein sein. Die Metrik, die für die Distanzen verwendet wird, darf kleine Änderungen nicht zu stark bewerten, und die Analyse sollte sich auf den Bereich der Daten beschränken, in dem die Formen der Transienten am stärksten voneinander abweichen.
Die auf der linken Seite dargestellten Cluster zeigen Datensätze mit „normalem" Verhalten. In Cluster 28 wurden 39 Transienten sortiert, deren Vorzeichen positiv ist, und in Cluster 31 werden 170 Transienten mit negativem Vorzeichen zusammengefaßt.
Die anderen beiden Bilder von Abbildung 5 zeigen Cluster mit ungewöhnlichem Verhalten. Die drei Transienten aus Cluster 13 kehren nach dem Schaltvorgang nicht vollständig auf das Nullniveau zurück. Cluster 2 ist ein Beispiel für einen Transienten mit einer außergewöhnlich

steilen Form. Er wird zu keinem anderen Transienten sortiert.

4 Zusammenfassung

Die Clusteranalyse eignet sich gut zur schnellen Sichtung großer Datenmengen. Sie kann aber nicht ohne Kenntnisse der Hintergründe als „Black-Box" verwendet werden. Methode, Metrik, Vorsortierung und Normierung müssen auf das jeweilige Problem angepaßt sein, damit die Clusteranalyse mit Erfolg eingesetzt werden kann.
Bei der Sortierung von unterschiedlichen Daten nach Form ist es besonders wichtig, nur den wirklich relevanten Teil der Daten zu betrachten und eine Metrik zu wählen, die genügend Spielraum läßt, um ähnliche aber nicht gleiche Transienten noch in ein Cluster zu sortieren.

5 Danksagung

Ich danke Anja Welker, die die Clusteranalyse für Problemstellungen in der Seismik benutzt und mich auf die Idee gebracht hat, diese Verfahren auch für elektromagnetische Zeitreihen zu verwenden.
Diese Arbeit wurde teilweise von der Europäischen Gemeinschaft unter der Projektnummer OG/0305/92/NL-UK unterstützt.

Literatur

Bigalke J., 1996, Untersuchungen zum nichtlinearen IP-Effekt zur Lokalisation ausgedehnter elektronenleitender Strukturen:16. Kolloquium Elektromagnetische Tiefenforschung, Deutsche Geophysikalische Gesellschaft, Protokollband, *in diesem Band*

Lance, G. H. und Williams, W. T., 1966: A General Theory of Classificatory Sorting Strategies I. Hierachical Systems: Computer Journal **9**, 373-380

Steinhausen D. und Langer L., 1977, Clusteranalyse Einführung in die Methoden und Verfahren der automatischen Klassifikation: Walter de GruYter, Berlin, New York

Strack, K. M., 1992, Exploration with deep transient electromagnetics: Elsevier Amsterdam, London,New York, Tokio

Sylvester, D., 1996, Neue LOTEM Messungen im Umfeld der KTB: 16. Kolloquium Elektromagnetische Tiefenforschung, Deutsche Geophysikalische Gesellschaft, Protokollband, *in diesem Band*