Density-Based Clustering in large Databases using Projections and Visualizations

Alexander Hinneburg
Institut für Informatik
Martin-Luther-Universität Halle-Wittenberg
hinneburg@informatik.uni-halle.de

1 Einleitung

Die Datenmengen, die in Computersystemen gespeichert werden, wachsen in aktuellen Anwendungsszenarien mit unvermindert großer Geschwindigkeit. Weil die Kapazität von Analysten begrenzt ist, sind automatische Methoden zur Extraktion von nützlichem Wissen aus großen Datenbanken sehr gefragt. Im Forschungsgebiet Data Mining ist diese Fragestellung ein Kernproblem und es wurden und werden viele verschiedene Ansätze vorgeschlagen, um unbekannte, interessante und nützliche Muster in großen Datenbanken zu entdecken.

Data Mining ist ein interdisziplinäres Forschungsgebiet, das Unterstützung aus den Bereichen Datenbanken, Statistik, Maschinellem Lernen, Visualisierung und vielen anderen erfahren hat. Aufgrund der starken Interdisziplinarität werden viele innovative Lösungen vorgeschlagenen, die mittels von Ansätzen aus verschiedenen Forschungsgebieten versuchen ähnliche Data Mining Aufgabenstellungen zu lösen. Jedoch, aufgrund der unterschiedlichen Forschungshintergründe und der verschiedenen verwendeten Paradigmen, ist ein Vergleich der Ergebnisse dieser Algorithmen schwierig.

Clusteranalyse ist neben dem Finden von Assoziations-Regeln und Klassifikation eine Basistechnik im Bereich Wissensentdeckung in Datenbanken. Ziel der Clusteranalyse ist es Objekte zu Gruppen zusammenzufassen, so daß ähnliche Objekte der selben Gruppe zugeordnet werden, Objekte aus verschiedenen Gruppen aber unähnlich zueinander sind. Clusteranalyse kann für unterschiedliche Aufgabenstellungen eingesetzt werden, u.a. zur Entdeckung natürlicher Klassen, zur Datenreduktion durch das Bestimmen repräsentativer Punkte oder zur Ausreißererkennung. Typische Anwendungen sind Kundensegmentierung, Dokument- oder Bildkategorisierung ebenso wie das Finden von Klassen in wissenschaftlichen Daten. Die bekannten Algorithmen setzen oft eine speziellere Clusterdefinition voraus, so daß häufig ein tieferes Verständnis des jeweiligen Algorithmus notwendig ist, um zu entscheiden ob er für eine gegebene Aufgabenstellung geeignet ist. Aus Anwendungssicht stellen sich zwei Problemfragen:

1. Wie wählt man einen geeigneten Algorithmus für eine spezielle Aufgabenstellung

zur Clusteranalyse aus?

2. Nach welchem Forschungsparadigma sollte der Algorithmus entworfen sein, um zu den Rahmenbedingungen der Anwendung zu passen?

Die erste Frage setzt sich mit Inhalt und Zweck der Clusteranalyse auseinander. Im Hinblick auf die Wahl des Algorithmus ist es wichtig zu wissen, ob z.B. das Finden natürlicher Klassen (mit Beziehungen zwischen den Attributen) oder eine möglichst verlustarme Datenreduktion gefragt ist. In vielen Forschungsartikeln wird eine solche Unterscheidung nicht explizit gemacht, da es Randfälle gibt, in denen die verschiedenen Aufgabenstellungen ähnlich sind.

Bei der zweiten Frage geht es darum, ob der von den Autoren einer Methode gewählte Kompromiß zwischen Effektivität und Ressourcenverbrauch zu der jeweiligen Anwendung paßt. Aufgrund der Heterogenität der vorgeschlagenen Ansätze, der oft sehr engen Verbindung der Ideen zur Beschleunigung der Clusterverfahren und der inhaltlichen Ausrichtung auf eine spezielle Aufgabenstellung ist eine davon unabhängige Wahl eines Algorithmus mit geeignetem Ressourcenverbrauch kaum möglich. Im ersten Teil der Arbeit wird neben der Identifizierung und Trennung der inhaltlichen Ausrichtungen der verschiedenen Clusterverfahren eine verfahrensunabhängige Methode zur Kontrolle des Ressourcenverbrauchs vorgeschlagen, das auf der statischen Methode der Dichteschätzung basiert.

Weitere Probleme ergeben sich bei der Behandlung von Daten, deren Objekte durch sehr viele Attribute (Dimensionen) beschrieben sind (20-500 Attribute pro Objekt). In solchen komplexen, hoch-dimensionalen Datenräumen läßt sich die Wahrscheinlichkeitsdichte nicht mehr statistisch signifikant schätzen [Si86, Sc92] und deshalb können keine Cluster gefunden werden. Dieses Problem haben nicht nur dichte-basierte Clusteralgorithmen, sondern alle anderen Methoden auch, aber durch die Dichteschätzung wird die schlechter werdende Qualität der Ergebnisse bei zunehmender Dimensionalität bemerkt. Im zweiten Kapitel der Dissertation wird ein neuer Algorithmus vorgeschlagen, der Cluster in niedrig-dimensionalen Projektionen von hoch-dimensionalen Datenräumen finden kann und die Cluster aus verschiedenen Projektionen zu einem Endergebnis zusammenfaßt.

Im Bereich Data Mining ist es sehr wichtig, daß die Ergebnisse von Menschen interpretiert, evaluiert und verstanden werden können. Erst dann kann man bei den gefundenen Mustern von entdecktem Wissen sprechen. Um dieses Aspekt zu unterstützen, wurde das Visualisierungssystem *HD-Eye* entwickelt, daß nicht nur erlaubt die Ergebnisse in graphischer Form zu präsentieren. Der hierverfolgte Ansatz besteht darin den Clusteranalyseprozess iterativ zu strukturieren. In jedem Iterationschritt wird das Clustermodell erweitert und der Anwender erhält schon in einem frühen Zwischenstadium visuelles Feedback und kann den Suchraum auf für die Anwendung relevante Bereiche eingrenzen. Das datenbank-gestützte System wurde schon am mehreren großen internationalen Datenbank-Konferenzen demonstriert [HKW02, HKW03b].

2 Clusteranalyse mit Hilfe von Primitiven

In der vorgelegten Dissertation wurde im ersten Teil ein Rahmen entwickelt, in dem vier verschiedene Aufgabenstellungen für Clusteranalyse identifiziert wurden, nämlich Datenreduktion, Finden natürlicher Klassen, Rauschfilterung und Ausreißererkennung. Für diese Aufgabenstellungen wurden effiziente Cluster-Primitive vorgeschlagen. In der Arbeit wird argumentiert, daß die Wahl der Aufgabenstellung von der inhaltlichen Verwendung des Ergebnisses abhängt und nicht durch automatische Verfahren aus den Daten abgeleitet werden kann.

Die Frage nach dem zu nutzenden Paradigma hängt stark mit der benötigten Effizienz des gewünschten Algorithmus zusammen. In der Literatur zur Clusteranalyse wurden für einzelne Aufgabenstellungen unter anderem folgende Ansätze benutzt: Bildkompression, mehrdimensionale Quadtrees, Gridfiles und neuronale Netze. In dieser Arbeit wurde Dichteschätzung als Grundlage der Clusteranalyse gewählt. Zum einen ist dieses Gebiet im Rahmen der angewandten Statistik gut erforscht und zum anderen wurde in der vorliegenden Promotionsarbeit gezeigt, daß die vorgeschlagenen Clustering-Primitive auf beliebigen Dichteschätzungsmethoden basieren können. Es wurde gezeigt, daß die entwickelte Trennung von Dichteschätzung und Clusteranalyse zu Algorithmen führt, die eine niedrigere oder gleichwertige Laufzeitkomplexität haben, als bisher bekannte Verfahren.

Ein weiterer Vorteil der Trennung in Dichteschätzung und Clusteranalyse ist, daß der datenaufwendige Teil der Dichteschätzung direkt in einem Datenbanksystem durchgeführt werden kann, ohne daß große Datenmengen zu einem Analyseanwendungsprogramm aus der Datenbank transportiert werden müssen. Dies verringert die Datenkommunikation erheblich. Verdeutlicht werden kann dies an folgendem Beispiel: gegeben seien 100000 zwei-dimensionale Datenpunkte in der Tabelle $data(x_1, x_2)$ und als Dichteschätzungsmethode wurde ein Histogramm mit 20 Intervallen pro Dimension gewählt. Anstatt die 100000 Tupel über eine ODBC-Schnittstelle zu dem Analyseanwendungsprogramm zu übertragen, kann das zwei-dimensionale Histogramm mit folgendem SQL-Statement direkt in der Datenbank berechnet werden:

```
SELECT floor((x1 - min1)/(max1 - min1) * 20),
       floor((x2 - min2)/(max2 - min2) * 20),
       COUNT(*) AS density
FROM data
GROUP BY floor((x1 - min1)/(max1 - min1) * 20),
         floor((x2 - min2)/(max2 - min2) * 20)
```

Als Resultat müssen nur die $20 \cdot 20 = 400$ Tupel des Histogramms übertragen werden. Auf diese Weise können auch mit handelsüblichen PCs mit weniger Systemressourcen (Hauptspeicher, CPU-Leistung) sehr große Datenmengen analysiert werden, da ein Großteil der Arbeit an den Datenbankserver delegiert werden kann. Der mit der Dichteschätzung einhergehende Informationsverlust ist für die Bestimmung der Cluster meist unerheblich, da diese aus größeren Bereichen bestehen und aus mehreren Histogrammzellen zusammengesetzt werden. Mit etwas aufwendigeren Dichteschätzungsmethoden, wie der k-Repäsentantenmethode oder Average Shifted Histograms (ASH), die ebenfalls in SQL formuliert

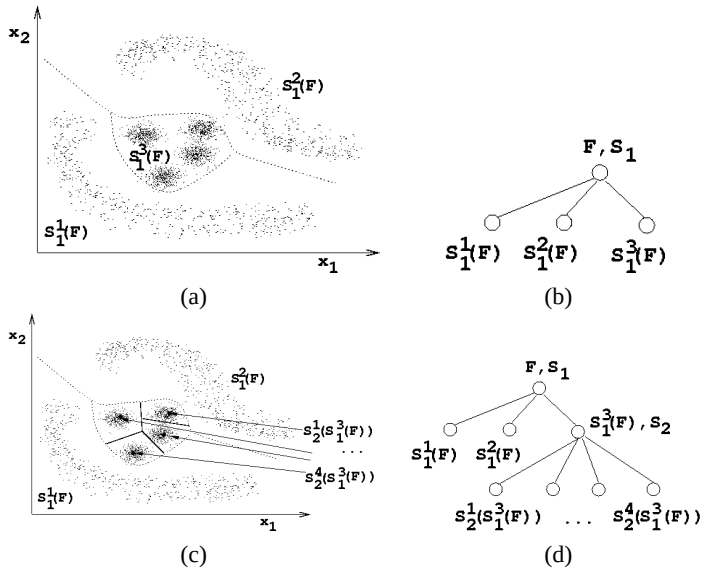


Abbildung 1: Beispiel für das hierarchisch rekursive Aufteilen des Datenraumes F in drei Gebiete mit beliebig geformten Clustern (a) und dem weiteren Separieren der vier normalverteilten Cluster in Gebiet $S_1^3(F)$ durch den k-Means ähnlichen Datenkompressionseparator S_2.

werden können, bleiben auch feinere Strukturen wie zum Beispiel Mikro-Cluster erhalten.

Die beiden in dieser Arbeit entwickelten Ansätze (Entwicklung von Cluster-Primitiven, Trennung von Dichteschätzung und Clusteranalyse) wurden in einem Rahmensystem integriert. Das System wurde so konzipiert, daß Cluster mittels sukzessiver Teilung der Daten gefunden werden. Für eine Teilung wird ein Cluster-Primitive (Separator) berechnet. Die entstehenden Teilmengen können dann von weiteren Separatoren zerlegt werden. Auf diese Art und Weise können komplexe, hierarchische Clustermodelle erzeugt werden. Eine Besonderheit bei dieser Vorgehensweise ist, daß für verschiedene Regionen im Datenraum verschiedene Primitive oder speziell angepaßte Parametrisierungen verwendet werden können. Dies wird in Abbildung 1 verdeutlicht, in der auf der ersten Stufe drei Gebiete mit beliebig geformten Clustern unterschieden wurden und auf der zweiten Stufe das Gebiet $S_1^3(F)$ nochmals in vier normal-verteilte Cluster durch einen k-Means ähnlichen Datenkompressionseparator aufgeteilt wurde.

Da im allgemeinen beliebig viele hierarchisch rekursive Unterteilungen erlaubt sind, kann die Zahl der Parameter des hierarchischen Cluster-Modells (Gesamtzahl aller Parameter der genutzten Separatoren) den Daten sinnvoll angepaßt werden.

3 Clusteranalyse in Projektionen

Die gegebenen Datenobjekte werden oft durch mehrdimensionale Vektoren beschrieben. Aus der Theorie der Dichteschätzung ist bekannt, daß mit steigender Dimensionalität die

Signifikanz der Ergebnisse abnimmt. Dieser Effekt (curse of dimensionality) läßt sich bei allen existierenden Algorithmen beobachten. Ein Ausweg ist Cluster in niedrig-dimensionalen Projektionen des hoch-dimensionalen Raumes zu suchen. Da die Anzahl der Projektionen sehr groß ist, werden die neu entwickelten effizienten Algorithmen aus Teil eins zur Analyse der Projektionen verwendet.

Das Teilgebiet Projected Clustering ist ein sehr neuer Ansatz, zu dem erst wenige Verfahren in der Literatur beschrieben sind. Der in der Dissertation erarbeitete Zugang zu Projected Clustering führt über Ähnlichkeitssuche. In [HAK00] wurde gezeigt, daß durch die Wahl einer geeigneten Teilmenge aus der Menge aller Attribute (Projektion), die zur Berechnung der Distanz zwischen zwei Objekten zugrunde gelegt wird, die Effektivität der Ähnlichkeitssuche gesteigert werden kann, gegenüber dem Fall, daß alle Attribute in die Distanzberechnung eingehen. Die optimale Wahl der Attribute für die Distanzberechnung ist von dem gewählten Anfrageobjekt (ein Punkt im mehrdimensionalen Raum) für die nächste Nachbarnanfrage abhängig.

Da Clusteranalyse stark vom dem gewählten Ähnlichkeitsmaß bzw. Distanzmaß abhängt, scheint es auch sinnvoll zu sein einen Cluster in einer Projektion, d.h. nur auf einer Teilmenge der Menge aller Attribute, zu definieren. Die Attribute in dieser Teilmenge sind die relevanten Attribute für den speziellen Cluster. Analog zur Ähnlichkeitssuche ist die geeignete Projektion auch von den Objekten im Cluster abhängig. In der vorliegenden Dissertation wurden die Probleme von existierenden Ansätzen aufgezeigt. Der CLIQUE Algorithmus [AGGR98], der als der geeigneteste Algorithmus für Data Mining identifiziert wurde, wurde in der Promotionsarbeit um zwei wesentliche Punkte erweitert:

1. Die nummerischen Daten wurden in Abhängigkeit der Datenverteilung diskretisiert. Der Original-Algorithmus zerlegte die Daten unabhängig von der Datenverteilung und konnte so in einigen Fällen nicht alle relevanten Attribute zu einem Cluster finden. Dies wird mit dem neuen Algorithmus vermieden.

2. Ähnliche Cluster, die in verschiedenen Projektionen definiert sind, werden zu einem Cluster zusammengefaßt. Dadurch können Cluster mit mehr als 6-8 relevanten Attributen gefunden, die CLIQUE nicht identifizieren konnte.

In der Promotionsarbeit wurde die Effektivität des neuen Algorithmus an Anwendungsbeispielen aus dem Bereich Bilddatenbanken und auf molekular-biologischen Daten demonstriert.

4 Visualisierung & Clusteranalyse

Der letzte Teil der Arbeit befaßt sich mit den Problemen der Anpassung und Parametrisierung der Algorithmen für gegebene Aufgabenstellungen, der Einbeziehung des Anwenders mit seinem Hintergrund- und Domänwissens in den Analyseprozess und der Evaluation der gefundenen Cluster. Wichtig ist bei allen drei Problemkomplexen die Einbeziehung des Anwenders. Oft gelingt es nur unvollkommen automatische Algorithmen an die

aktuelle Aufgabe mittels Parameter anzupassen. Ebenfalls ist es wichtig dem Anwender einen groben Überblick über Struktur der Daten zu vermitteln, so daß sie/er die Daten mit Hintergrundwissen verknüpfen kann. Dafür eignen sich insbesondere Visualisierungen, die eine große Menge an Daten dem Anwender schnell nahe bringen können. Im letzten Teil der Arbeit wurden neue Visualisierungen entwickelt, die Eigenschaften von Projektionen hochdimensionaler Räume mittels Icons und Pixel-orientierter Grafiken darstellen. In Kombination mit effizienten automatischen Algorithmen kann der Anwender erstes einen besseren Eindruck von der Ausgangsdaten erhalten und zweitens die Effektivität der Clusteranalyse beträchtlich steigern, weil sie/er die automatischen Verfahren besser entsprechend der Daten und seinem Hintergrundwissen parametrisieren kann. So können mit einem System, das automatische Verfahren mit Visualisierungen verknüpft, Cluster gefunden werden, die mit einfachen automatischen Verfahren nicht entdeckt werden. Ein Prototyp eines solches Systems, *HD-Eye*, wurde im Rahmen der Promotion entwickelt. Der Prototyp wurde auf großen internationalen Tagungen (SIGMOD 2002 und ICDE 2003) der Öffentlichkeit demonstriert [HKW02, HKW03b]. Abbildung 2 zeigt zwei Screenshots unseres Prototypen.

Das visuelle Clusteranalyse-System *HD-Eye* unterstützt automatische Clusterungsverfahren wie *DENCLUE* [HK98] bzw. *OptiGrid* [HK99], LBG und DBSCAN. Die ersten Ergebnisse wurden in [HWK99] veröffentlicht, die zweite Version von *HD-Eye* wurde in [HKW03a] beschrieben. Als besondere Muster in den Daten werden lokale Maxima in der Wahrscheinlichkeitsdichtefunktion hervorgehoben, da in diesen Bereichen die Daten besonders stark geclustert sind. Als Visualisierungstechniken werden zum einen Color-Density Plots von 1D und 2D Projektionen (siehe Abb. 3 und 4) der Daten und zum anderen Icon-basierte Visualisierungen der Maxima der Dichtefunktion (Abb. 6) verwendet. Bei der Icon-basierten Visualisierung der Maxima werden insbesondere die Relevanz (Anzahl der Datenpunke des Maximums) sowie die Separiertheit der Maxima (Verhältnis zwischen maximaler Dichte und Dichte am Rand) dargestellt (Abb. 5).

HD-Eye unterstützt visuell auch die Parameterwahl für die implementierten automatischen Algorithmen. Wichtige Parameter sind die Rauschschwelle ξ oder die Anzahl der Repräsentanten zum Beschreiben der Cluster. In Abbildung 7 wird die Anpassung an zwei Beispielen demonstriert.

Insgesamt erlaubt die Visualisierung dem Benutzer, die Struktur der Daten besser zu verstehen sowie interaktiv die besten Projektionen auszuwählen, wodurch die Effektivität der automatischen Clusterungsverfahren erheblich gesteigert wird.

5 Zusammenfassung

Es wurde ein Rahmensystem für Clusteranalyse entwickelt, daß Cluster-Primitive für verschiedene Aufgabenstellungen bereit hält. Alle Cluster-Primitive basieren auf Dichteschätzung, die von der eigentlichen Clusteranalyse getrennt wurde. Diese Trennung führte zu Algorithmen mit besser Laufzeitkomplexität. Um hoch-dimensionale Daten zu bearbeiten wurde ein neuer Algorithmus vorgeschlagen, der Cluster in verschiedenen Projektionen

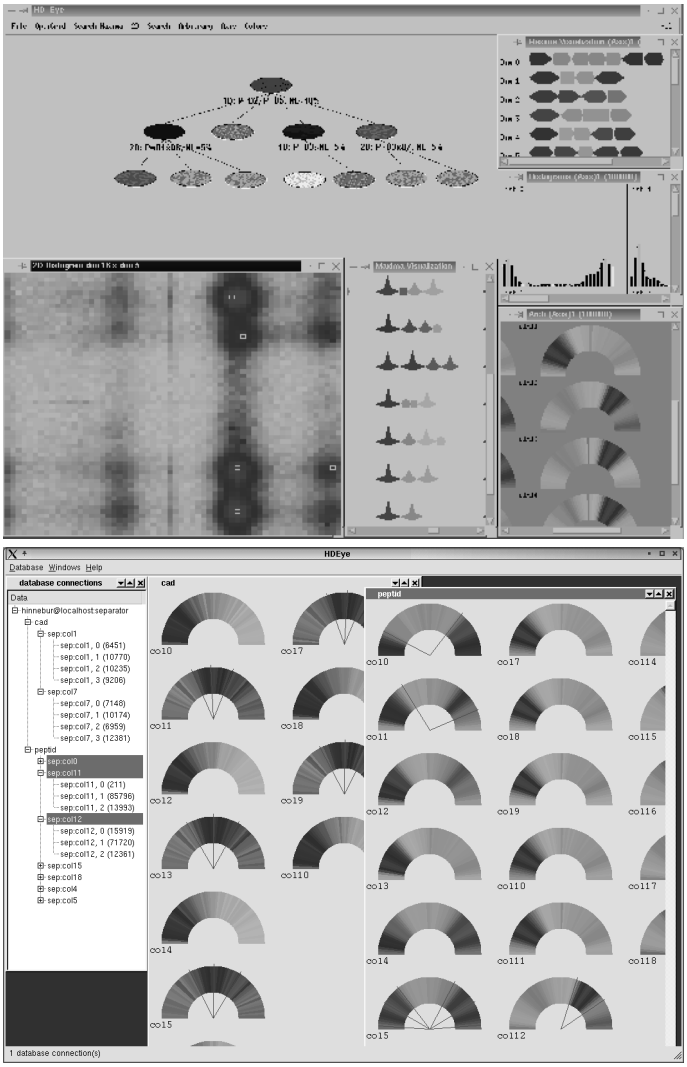
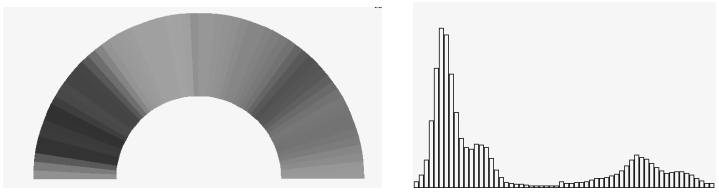


Abbildung 2: HD-Eye Screenshot Version 1 and Version 2, Erklärung der Teilfenster in der oberen Abbildung im Uhrzeigersinn von Oben: Separator Baum, Icon Repräsentation von 1D Projektionen, 1D Projektion-Histogramm, 1D Dichte Diagramm, Icon Repräsentation für multi dimensionale Projektionen and 2D Dichte Diagramme.



(a) Color (b) Histogramm

Abbildung 3: Beispiel für ein-dimensionale Color-Density Plots

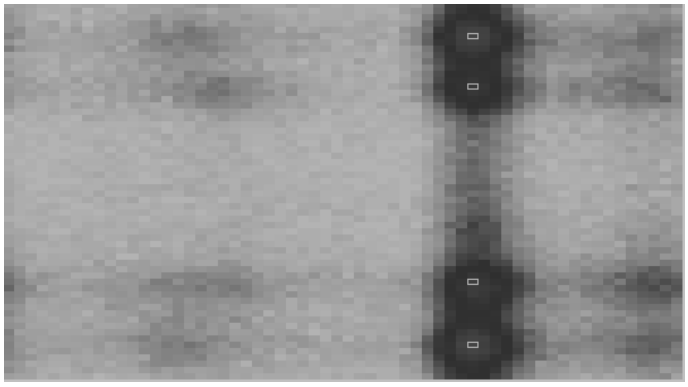
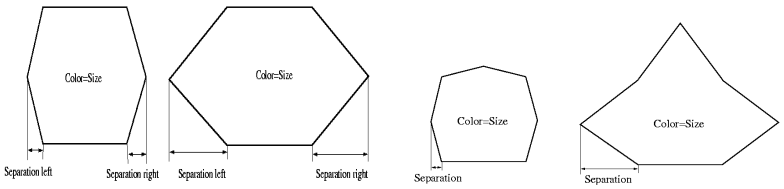


Abbildung 4: Beispiel für einen zwei-dimensionalen Color-Density Plot

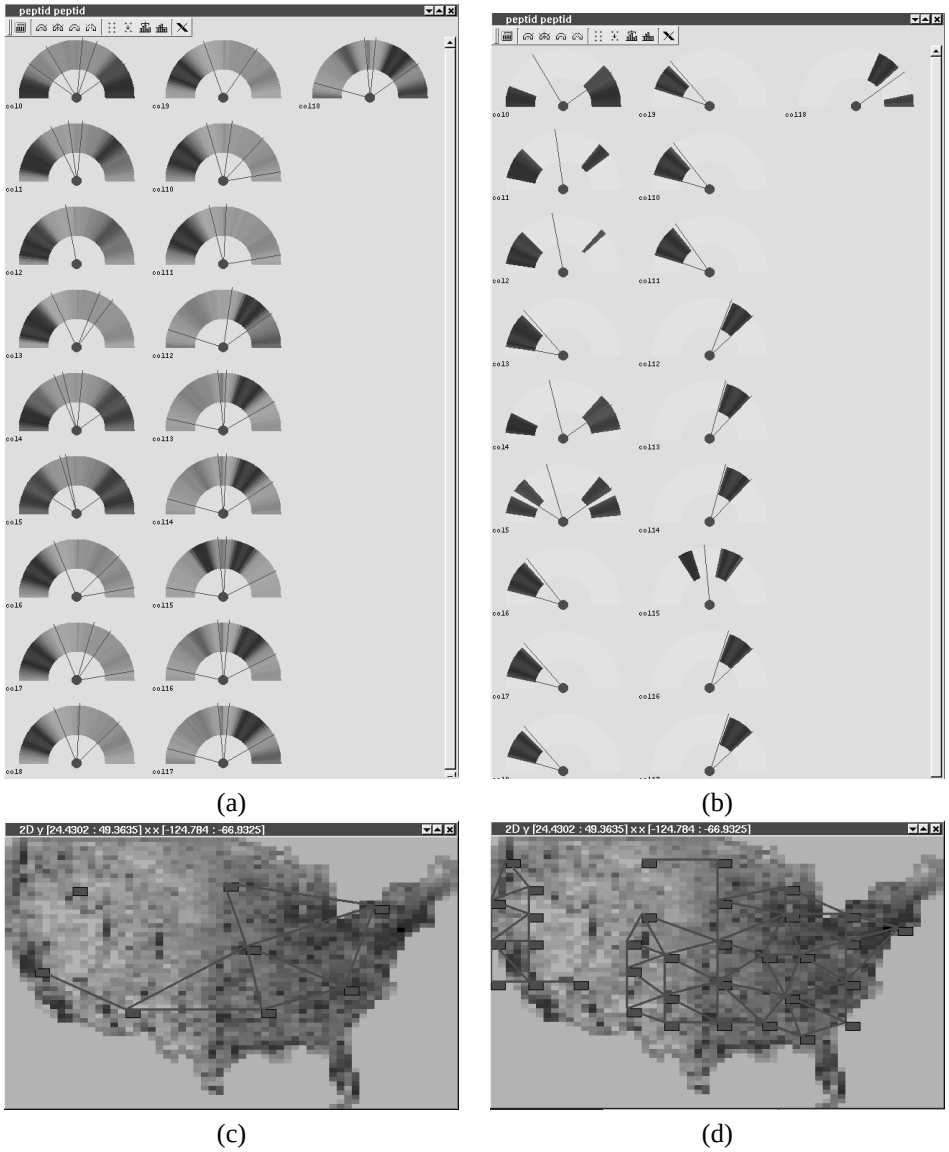


(a) 1 dimensional (b) multidimensional

Abbildung 5: Struktur der Icons

(a) ein-dimensional (b) mehr-dimensional

Abbildung 6: Beispiele für Icons passend zu den vorhergehenden Color-Density Plot in Abb.3 und 4



(a) (b)

(c) (d)

Abbildung 7: (a) zeigt Color-Density Plots von molekular-biologischen Daten mit den separarierenden Minima für die Rauschschwelle $\xi = 0$. Aufgrund der Visulisierungen erhöht der Anwender die Rauschschwelle auf 2%. Teil(b) zeigt die veränderten Density-Plots, wobei die Intervalle mit einer Dichte unterhalb der Rauschschwelle gelb gezeichnet sind. Mit Hilfe der Rauschschwelle werden Trennpunkte entfernt, die durch leichte Schwankungen in der Datenverteilung verursacht werden. Die Teile (c,d) zeigen wie eine größere Menge von Repräsentanten die Approximationsqualität der Cluster verbessert. In dem Beispiel werden in den Daten des US Census Büros die dichten geclusterten Gebiete der West- und Ostküste getrennt.

des hoch-dimensionalen Datenraumes finden kann. Der neue Algorithmus kann Cluster finden, die von anderen bekannten Verfahren nicht gefunden werden. Zum Abschluß wurde das *HD-Eye*-System entwickelt, das automatische Verfahren mit Visualisierungstechniken verknüpft, um dem Nutzer eine bessere Grundlage für seine Entscheidungen zu liefern und um das Verständnis und die Einschätzung der Ergebnisse zu erleichtern. In zukünftigen Arbeiten kann der Algorithmus um das Finden von Clustern mit abhängigen Attributen erweitert werden. In diesem Rahmen gibt es auch Potential zur Entwicklung neuer Visualisierungstechniken. Ebenso können Verfahren für nominale Daten (im Gegensatz zu den hier genutzten nummerischen Daten) untersucht werden.

Literatur

[AGGR98] Agrawal, R., Gehrke, J., Gunopulos, D., und Raghavan, P.: Automatic subspace clustering of high dimensional data for data mining applications. In: *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, 1998, Seattle, Washington, USA*. S. 94–105. ACM Press. 1998.

[HAK00] Hinneburg, A., Aggarwal, C. C., und Keim, D. A.: What is the nearest neighbor in high dimensional spaces? In: *VLDB'2000, Proceedings of 26th International Conference on Very Large Data Bases, Cairo, Egypt*. S. 506–515. Morgan Kaufmann. 2000.

[HK98] Hinneburg, A. und Keim, D.: An efficient approach to clustering in large multimedia databases with noise. In: *KDD'98, Proc. of the 4th Int. Conf. on Knowledge Discovery and Data Mining*. S. 58–65. AAAI Press. 1998.

[HK99] Hinneburg, A. und Keim, D. A.: Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. In: *VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases, September 7-10, 1999, Edinburgh, Scotland, UK*. S. 506–517. Morgan Kaufmann. 1999.

[HKW02] Hinneburg, A., Keim, D. A., und Wawryniuk, M.: Hdeye: Visual mining of high-dimensional data (demo). In: *SIGMOD 2002, Proceedings ACM SIGMOD International Conference on Management of Data, June 3-6, 2002, USA*. ACM Press. 2002.

[HKW03a] Hinneburg, A., Keim, D. A., und Wawryniuk, M.: Using projections to visually cluster high–dimensional data. *IEEE Computing in Science & Engineering*. 5(2):14–25. 2003.

[HKW03b] Hinneburg, A., Keim, D. A., und Wawryniuk, M.: Hdeye: Visual mining of high-dimensional data (demo). In: *ICDE 2003, Proceedings of the 19th International Conference on Data Engineering, ICDE, India*. IEEE Press. 2003.

[HWK99] Hinneburg, A., Wawryniuk, M., und Keim, D. A.: Hdeye: Visual mining of high-dimensional data. *Computer Graphics & Applications Journal*. 19(5):22–31. September/October 1999.

[Sc92] Scott, D.: *Multivariate Density Estimation*. Wiley and Sons. 1992.

[Si86] Silverman, B. W.: *Density Estimation for Statistics and Data Analysis*. Chapman & Hall. 1986.