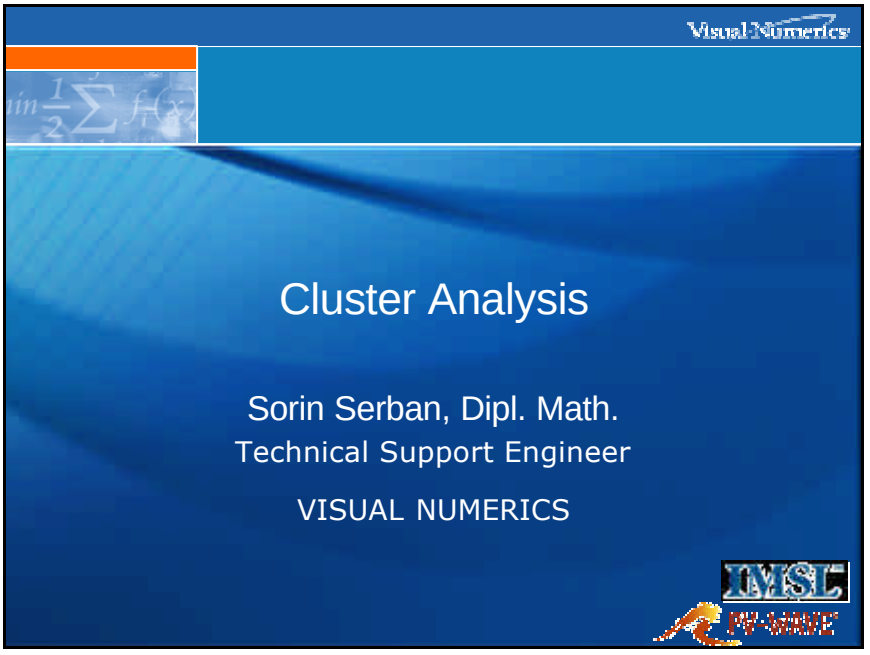
Visual Numerics
JavaForum 2007
IMSL
PV-WAVE


Visual Numerics
Cluster Analysis
Sorin Serban, Dipl. Math.
Technical Support Engineer
VISUAL NUMERICS
IMSL
PV-WAVE

Visual Numerics

AGENDA

- Einleitung (Visual Numerics)
- K-means Cluster
- Hierarchische Cluster
- Beispiele
- Fragen und Antworten

IMSL PV-WAVE

Visual Numerics

Visual Numerics Global Footprint

IMSL PV-WAVE

Visual Numerics

Products & Services

IMSL

PV-WAVE

- Embeddable numerical analysis
  - Available for C/C++, Fortran, .NET, and **Java (our focus today)**
- Sophisticated Visual Data Analysis
  - Development environment including time-series and web-based functionality
- Expert Professional Services for customized algorithm and application development
- Worldwide presence for local customer support

Visual Numerics

IMSL® Library Techniques

Mathematics

- Basic Types
- Linear Algebra
- Eigensystems
- Interpolation & Approximation
- Quadrature
- Differential Equations
- Nonlinear Equations
- Numerical Optimization
- Special Functions

Statistics

- Basic Statistics
- Time Series & Forecasting
- Nonparametric Tests
- Variances, & Covariances
- Correlations
- Regression
- Analysis of Variance
- Transforms
- Goodness of Fit
- Distribution Functions
- Random Number Generation

Visual Numerics

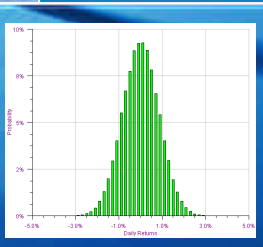
JMSL™ Numerical Library for Java Applications

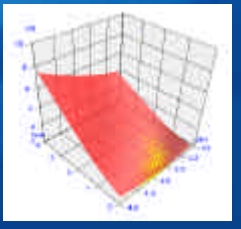
- Complete collection of mathematical, statistical and charting classes
- Written in 100% Java™
- Based upon the “gold standard” IMSL™ Library of algorithms
- Complete, robust, fully supported, documented

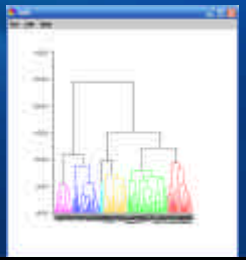
IMSL
PV-WAVE

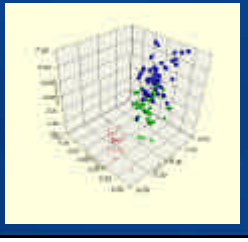
Visual Numerics

2D and 3D Charting (JMSL)









IMSL
PV-WAVE

Visual Numerics

Clusteranalyse

Grundlegende Fragen:

1. Was ist ein Cluster?
2. Was versteht man unter Clusteranalyse?
3. Wofür die Clusteranalyse?

IMSL PV-WAVE

Visual Numerics

Clusteranalyse

1. **Cluster** werden als Gruppe von Objekten (Anhäufung von Punkten, die in der Regel in Form von Vektoren gespeichert werden) definiert, die in Bezug auf einen berechneten Schwerpunkt eine minimale Abstandssumme haben.

IMSL PV-WAVE

Clusteranalyse

2. Unter **Clusteranalyse** versteht man ein strukturentdeckendes, multivariates Analyseverfahren zur Ermittlung von Gruppen (Clustern) von Objekten, deren Eigenschaften oder ihre Ausprägungen bestimmte Ähnlichkeiten oder Unähnlichkeiten aufweisen.

Clusteranalyse

3. Angepasste Verfahren der Clusteranalyse lassen sich zur

- automatischen Klassifikation,
- Erkennung von Mustern in der Bildverarbeitung,
- oder zum Data-Mining

einsetzen.

Visual Numerics

Clusteranalyse

Clusterverfahren

I. Partitionierende Verfahren
   a. k-means Algorithmus
II. Hierarchische Verfahren
   a. Anhäufende Algorithmen (nähere Betrachtung)
   b. Teilende Algorithmen

Visual Numerics

Clusteranalyse

1. Ausgehend von einem oder mehreren Mittelpunkten werden, über eine Abstandsfunktion *d,* (zufällig gewählte) Punkte mit den kleinsten Abständen zu den Mittelpunkten, zu Clustern zusammengefasst.

Clusteranalyse

II. Ausgehend von vielen Clustern, die nur einen Punkt enthalten, werden in jedem Schritt ähnliche Cluster zu größeren Clustermengen vereinigt (Umkehrung möglich).

Visual Numerics

Clusteranalyse

a. K-Means Algorithmus

- ***Initialisierung:*** (Zufällige) Auswahl von k Clusterzentren
- ***Zuordnung:*** Jedes Objekt wird dem ihm am nächsten liegenden Clusterzentrum zugeordnet
- ***Neuberechnung:*** Es werden für jedes Cluster die Clusterzentren neu berechnet
- ***Wiederholung:*** Falls sich nun die Zuordnung der Objekte ändert, weiter mit Schritt 2, ansonsten Abbruch

Clusteranalyse

Die Hauptidee des Algorithmus ist es, nach einer K-Partition mit einer lokal optimierten Summe der Euklidischen Entfernung zu suchen, indem man Punkte von einem Cluster zum anderen verschiebt.

$$D = \sum_{i=1}^{K} \sum_{j=1}^{p} \sum_{m=1}^{n_i} f_{n_{im}} w_{n_{im}} d_{n_{im},j} \left(x_{n_{im},j} - x_{i,j} \right)^2$$

Clusteranalyse

Eingabe = Matrix von *M* Punkten in *N* Dimensionen.

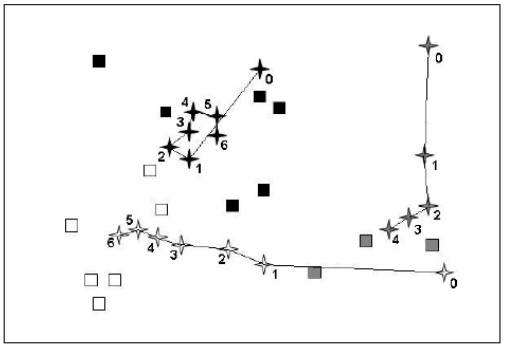
N(C) = bezeichnet die Anzahl der Punkte im Cluster *C*.

D(I,Z(C)) = Euklidische Entfernung zwischen dem Punkt *I* und dem Zentrum *Z* von Cluster *C*.

Clusteranalyse

1. Initialisiere Cluster *C(j), j=1,…,k* und berechne für jedes Cluster das Zentrum *Z(j) j=1,…,k.*
2. Für jeden Punkt *I(i), i=1,…,M* finde das nächstgelegene Zentrum *Z(m)* und füge *I* dem Cluster *C(m)* zu:
   - Berechne *D(I(i), Z(j))* für alle j=*1,..,k*
   - Bestimme Zentrum *Z(m)* so, dass *D(I(i), Z(m))* minimal ist (bei mehreren Zentren entscheide zufällig)
   - Füge Punkt *I* dem Cluster *C(m)* hinzu.

Clusteranalyse



Clusteranalyse

Konstruktor:

- public ClusterKMeans(double[][] x, double[][] cs)

Methode:

- final public double[][] compute() throws ClusterKMeans.NoConvergenceException, ClusterKMeans.ClusterNoPointsException
- public int[] getClusterCounts()
- public int[] getClusterMembership()
- public double[] getClusterSSQ()

Clusteranalyse

- public void setFrequencies(double[] frequencies) throws ClusterKMeans.NonnegativeFreqException
- public void setMaxIterations(int iterations)
- public void setWeights(double[] weights) throws ClusterKMeans.NonnegativeWeightException

Beispiel 1

Visual Numerics

Clusteranalyse

b. Anhäufender Algorithmus

- ***Initialisierung:*** Jedes Objekt wir als ein eigenes Cluster mit einem Element aufgefasst
- ***Zuordnung:*** Die jeweils einander am nächsten stehenden Cluster werden zu einem Cluster zusammengefasst
- ***Wiederholung:*** Falls nicht alle Cluster eine bestimmte Distanz zueinander haben oder keine genügend kleine Zahl von Clustern ermittelt worden ist, weiter mit Schritt 2, ansonsten Abbruch

IMSL PV-WAVE

Visual Numerics

Clusteranalyse

Die Hauptidee der Hierarchische Methoden – auch Linkage Methoden – ist, den Abstand von Punkten eines Clusters mit den Punkten eines anderen Clusters zu betrachten.

Folgende Linkage Verfahren werden unterschieden:

- ***Single Linkage***

  Minimaler Abstand zweier Punkte
- ***Complete Linkage***

  Maximaler Abstand zweier Punkte

IMSL PV-WAVE

Clusteranalyse

- ***Average Linkage***
  Durchschnittlicher Abstand der Punkte zweier Cluster
- ***Centroid***
  Abstand der Cluster-Zentren
- ***Ward's Methode***
  Betrachte die SSQ bzgl. der Cluster-Zentren

Clusteranalyse

Entfernungsmethoden

Konstruktor:

- public Dissimilarities(double[][] x, int distanceMethod, int distanceScale, int iRow, int[] indexArray) throws Dissimilarities.ScaleFactorZeroException, Dissimilarities.ZeroNormException, Dissimilarities.NoPositiveVarianceException

Visual Numerics

Clusteranalyse

Entfernungsmethoden

distanceMethod	Metric
0	Euclidean distance (L_2norm)
1	Sum of the absolute differences (L_1 norm)
2	Maximum difference (L_∞ norm)
3	Mahalanobis distance
4	Absolute value of the cosine of the angle between the vectors
5	Angle in radians (0, pi) between the lines through the origin defined by the vectors
6	Correlation coefficient
7	Absolute value of the correlation coefficient
8	Number of exact matches, where $x_i = y_i$.

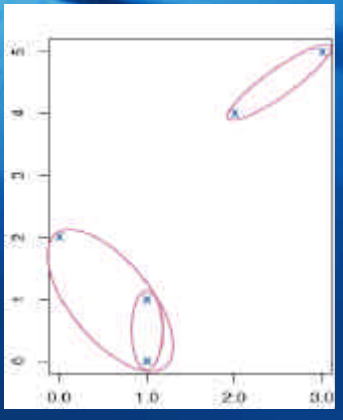
IMSL PV-WAVE

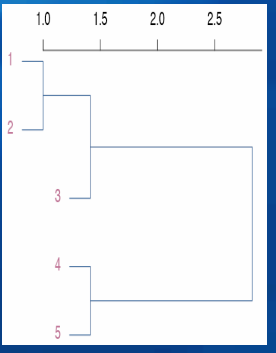
Visual Numerics

Clusteranalyse

Daten:

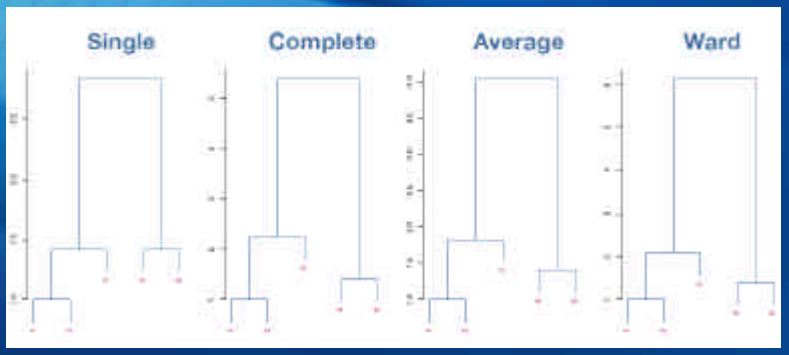
X1	X2
1	1
1	0
0	2
2	4
3	5





IMSL PV-WAVE

Clusteranalyse



Clusteranalyse

Konstruktor:

- public ClusterHierarchical(double[][] dist, int method, int transform)

Methoden:

- final public int[] getClusterLeftSons()
- final public double[] getClusterLevel()
- final public int[] getClusterMembership(int nClusters)
- final public int[] getClusterRightSons()
- final public int[] getObsPerCluster(int nClusters)

Beispiel 2

Visual Numerics

FRAGEN UND ANTWORTEN

Sorin Serban

Technical Support Engineer

serban@visual-numerics.de

+49 (0)711-67400-262

IMSL

PV-WAVE