Ludwig Maximilians Universität München
Institut für Informatik
Lehr- und Forschungseinheit für Datenbanksysteme

Skript zur Vorlesung

Knowledge Discovery in Databases

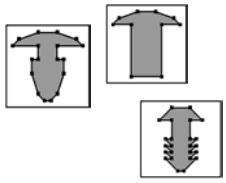
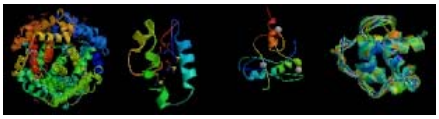
im Wintersemester 2007/2008

Kapitel 2: Merkmalsräume

http://www.dbs.ifi.lmu.de/Lehre/KDD

Merkmalsräume

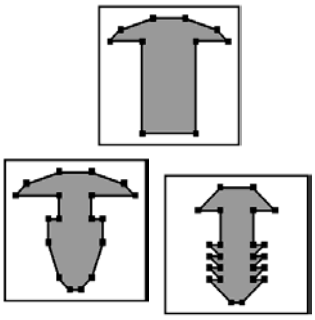
- Motivation:
  - Zentrales Konzept beim Data Mining: Ähnlichkeit von Datenbankobjekten
    - Clustering: Zusammenfassen *ähnlicher* Objekte in Gruppen
    - Klassifikation: Zuordnung von Objekten zu einer Klasse *ähnlicher* Objekte
  - Definition einer geeigneten Distanzfunktion auf Datenbankobjekten nicht immer einfach (besonders in Nicht-Standard-Datenbanken)
    - Bilder
    - CAD-Objekte
    - Proteine
    - Textdokumente
    - Polygonzüge (GIS)
    - etc.

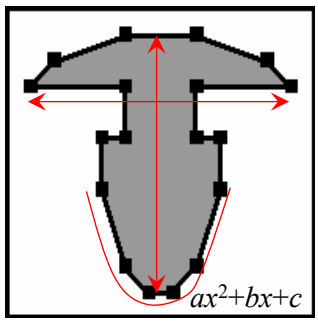





Merkmale („Features“ von Objekten)

- Oft sind die betrachteten Objekte komplex
- Eine Aufgabe des KDD-Experten ist dann, geeignete Merkmale (*Features*) zu definieren bzw. auszuwählen, die für die Unterscheidung (Klassifikation, Ähnlichkeit) der Objekte relevant sind.

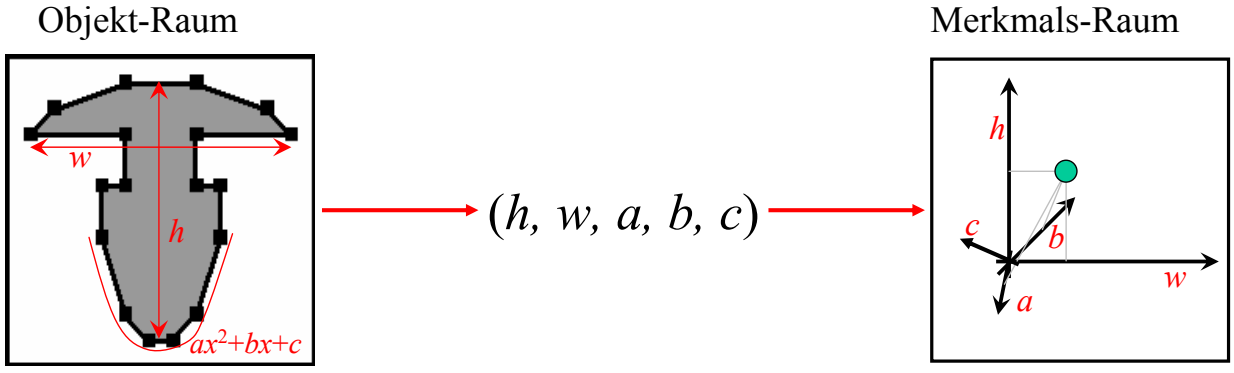
Beispiel: CAD-Zeichnungen:





Mögliche Merkmale:

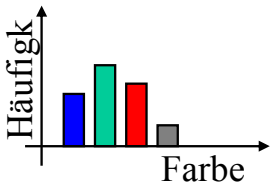
Beispiel: CAD-Zeichnungen (cont.)



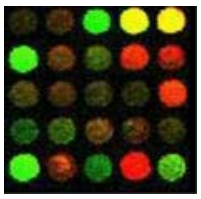
- Im Kontext von statistischen Betrachtungen werden die Merkmale häufig auch als *Variablen* bezeichnet
- Die ausgewählten Merkmale werden zu Merkmals-Vektoren (*Feature Vector*) zusammengefasst
- Der Merkmalsraum ist häufig hochdimensional (im Beispiel 5-dim.)

Merkmale

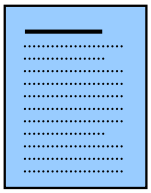
Bilddatenbanken:
Farbhistogramme



Gen-Datenbanken:
Expressionslevel



Text-Datenbanken:
Begriffs-Häufigkeiten



Data	25
Mining	15
Feature	12
Object	7
...	

Der Feature-Ansatz ermöglicht einheitliche Behandlung von Objekten verschiedenster Anwendungsklassen

Merkmale

Skalen-Niveaus von Merkmalen

Nominal (kategorisch)

Charakteristik:

Nur feststellbar, ob der Wert gleich oder verschieden ist. Keine Richtung (besser, schlechter) und kein Abstand. Merkmale mit nur zwei Werten nennt man *dichotom*

Beispiele:

Geschlecht (dichotom)
Augenfarbe
Gesund/krank (dichotom)

Ordinal

Charakteristik:

Es existiert eine Ordnungsrelation (besser/schlechter) zwischen den Kategorien, aber kein einheitlicher Abstand

Beispiele:

Schulnote (metrisch?)
Güteklasse
Altersklasse

Metrisch

Charakteristik:

Sowohl Differenzen als auch Verhältnisse zwischen den Werten sind aussagekräftig. Die Werte können diskret oder stetig sein.

Beispiele:

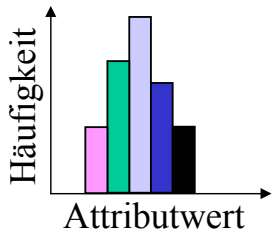
Gewicht (stetig)
Verkaufszahl (diskret)
Alter (stetig oder diskret)

Univariate Deskription von Merkmalen

Sei $x_1,\ldots,x_n$ eine Stichprobe eines Merkmals X.

- Absolute Häufigkeit: Für jeden Wert a ist h(a) die Anzahl des Auftretens in der Stichprobe
- Relative Häufigkeit: $f(a) = h(a) / n$

- Histogramm:



Die folgenden Maße sind nur für metrische Merkmale sinnvoll:

- Arithmetisches Mittel: $\mu = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$
- Median: *Das mittlere Element bei aufst. Sortierung*
- Varianz: $VAR(X) = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$
- Standardabweichung: $\sigma = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$

Multivariate Deskription von Merkmalen

- Kontingenztabelle
  - für kategorische Merkmale X und Y
  - repräsentiert für zwei Merkmale X und Y die absolute Häufigkeit h_{ik} jeder Kombination (x_i, y_k) und alle Randhäufigkeiten $h_{.k}$ und $h_{i.}$ von X und Y

	MittelfristigeArbeitslosigkeit	Langfristige Arbeitslosigkeit	
Keine Ausbildung	19	18	37
Lehre	43	20	63
	62	38	100

  - Wie sollten die relativen Häufigkeiten verteilt sein, wenn die beiden Merkmale keinerlei Abhängigkeit besitzen? $\frac{h_{ik}}{n} = \frac{h_{i.}}{n} \cdot \frac{h_{.k}}{n}$
  - χ^2-*Koeffizient*
    Differenz zwischen dem bei Unabhängigkeit erwarteten und dem tatsächlich beobachteten Wert von h_{ij} (Maß für die Stärke der Abhängigkeit)

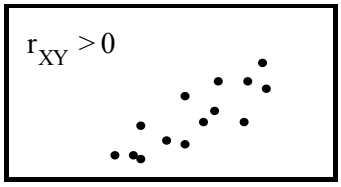
Multivariate Deskription von Merkmalen

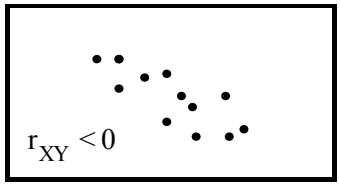
Korrelationskoeffizient

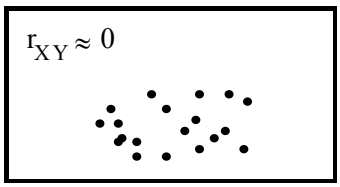
- für numerische Merkmale X und Y
- wie stark sind die Abweichungen vom jeweiligen Mittelwert korreliert?

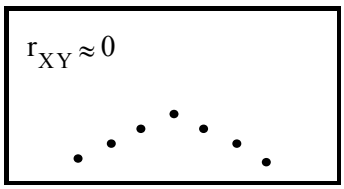
$$r_{XY} = \sum_{i=1}^{n} \frac{(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

- Beispiele









Räume und Distanzfunktionen

- Merkmalsraum (Featureraum)
  - Intuitiv: ein Wertebereich/Domain mit Distanzfunktion
  - Formal: Featureraum **F** = (*Dom*, *dist*)

    Dom ist eine (geordnete) Menge von Merkmalen (Features)

    $dist : Dom \times Dom \rightarrow \mathbb{R}$ ist eine totale (Distanz)-Funktion mit den folgenden Eigenschaften
    - $\forall p,q \in Dom, p \neq q : dist(p,q) > 0$
    - $\forall o \in Dom : dist(o,o) = 0$

      } Reflexivität
    - $\forall p,q \in Dom : dist(p,q) = dist(q,p)$ Symmetrie

Räume und Distanzfunktionen

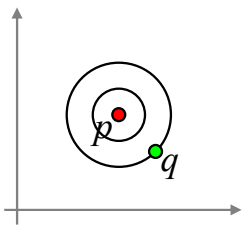
- Metrischer Raum
  - Formal: Metrischer Raum **M** = (*Dom*, *dist*) mit den folgenden Eigenschaften
    - **M** ist ein Featureraum
    - $\forall o,p,q \in Dom : dist(o,p) \leq dist(o,q) + dist(q,p)$ Dreiecksungleichung
- Wichtigstes Beispiel: Euklidischer Vektorraum
  - Formal: Euklidischer Vektorraum **E** = (*Dom*, *dist*) mit
    - (*Dom*, *dist*) ist ein metrischer Raum
    - $Dom = \mathbb{R}^d$
- Sprechweise:
  - Euklidischer Vektorraum = „Featureraum“
  - Vektoren (Objekte im Eulidischen Featureraum) = „Featurevektoren“
  - Die *d* Dimensionen des Vektorraums = „Features“

Räume und Distanzfunktionen

- Ähnlichkeit von Feature Vektoren (Euklidische Vektoren)

Euklidische Norm (L_2):

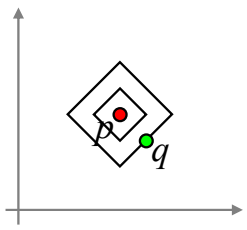
$dist_1 = ((p_1-q_1)^2+(p_2-q_2)^2+...)^{1/2}$



Natürlichstes Distanzmaß

Manhattan-Norm (L_1):

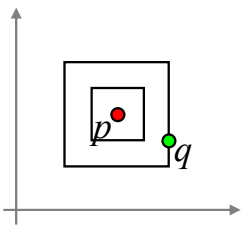
$dist_2 = |p_1-q_1|+|p_2-q_2|+...$



Die Unähnlichkeiten der einzelnen Merkmale werden direkt addiert

Maximums-Norm (L_∞):

$dist_\infty = \max\{|p_1-q_1|, |p_2-q_2|,...\}$



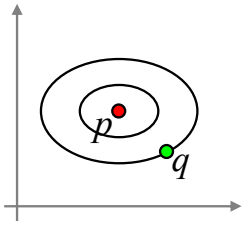
Die Unähnlichkeit des am wenigsten ähnlichen Merkmals zählt

Verallgemeinerung L_p-Abstandsmaß: $dist_p = (|p_1-q_1|^p + |p_2-q_2|^p + ...)^{1/p}$

Räume und Distanzfunktionen

Gewichtete Euklidische Norm:

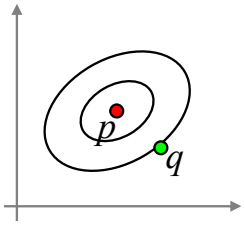
$dist = (w_1(p_1-q_1)^2 + w_2(p_2-q_2)^2+...)^{1/2}$



Häufig sind die Wertebereiche der Merkmale deutlich unterschiedlich.
Beispiel: Merkmal $M_1 \in [0.01 .. 0.05]$
Merkmal $M_2 \in [3.1 .. 22.2]$
Damit M_1 überhaupt berücksichtigt wird, muss es höher gewichtet werden

Quadratische Form:

$dist = ((p-q)\,\mathbf{M}\,(p-q)^{\mathbf{T}})^{1/2}$



Bei den bisherigen Ähnlichkeitsmaßen werden die Merkmale nur getrennt gewichtet.
Besonders bei Farbhistogrammen müssen auch *verschiedene* Merkmale gemeinsam gewichtet werden.

Statt mit Distanzmaßen, die die Unähnlichkeit zweier Objekte messen, arbeitet man manchmal auch mit positiven Ähnlichkeitsmaßen

Räume und Distanzfunktionen

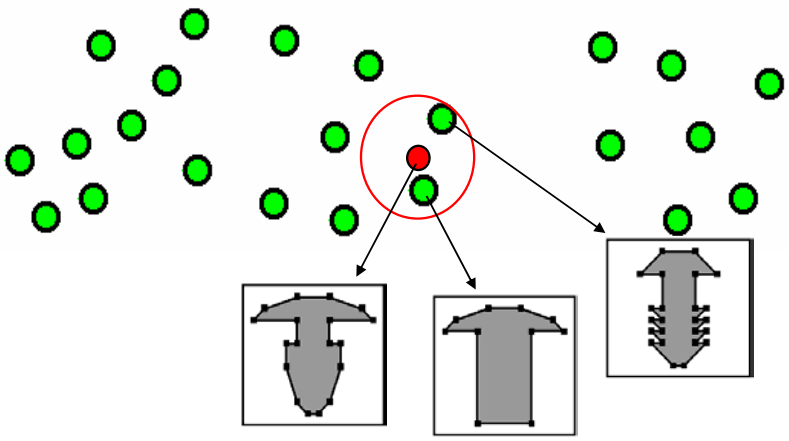
- Spezifiziere Anfrage-Objekt $q \in DB$ und…
  - … suche ähnliche Objekte – Range-Query (Radius ε)

$$RQ(q,\varepsilon) = \{ o \in DB \mid dist(q,o) \leq \varepsilon \}$$

  - … suche die k ähnlichsten Objekte – Nearest Neighbor

$NN(q,k) \subseteq DB$ mit mind. k Objekten, sodass

$$\forall o \in NN(q,k),\ p \in DB\text{-}NN(q,k) : dist(q,o) < dist(q,p)$$

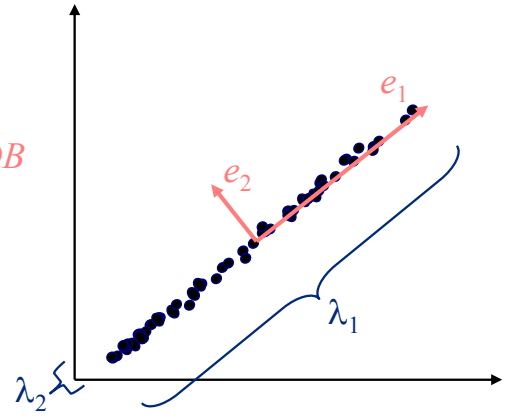


Räume und Distanzfunktionen

- Deskription von Featurevektoren
  - Gegeben: Menge *DB* von Featurevektoren
  - Zentroid (Centroid, vgl. Arithmetisches Mittel): $\mu_{DB} = \frac{1}{DB} \cdot \sum_{o \in DB} o$
    - Achtung: bei allgem. Metrischen Räumen muss Centroid nicht notwendigerweise existieren!!!
  - Medoid m_{DB}:
    - Der Featurevektor, der am nächsten zum Centroiden gelegen ist (die kleinste Distanz zum Zentroiden hat)
    - Bei allgem. Metrischen Räumen: Objekt mit dem kleinsten durschnittlichen Abstand zu allen anderen Objekten aus *DB*
  - Varianz (der Distanzen): $Var_{DB} = \frac{1}{DB} \cdot \sum_{o \in DB} dist(o, \mu_{DB})$
  - Standardabweichung analog

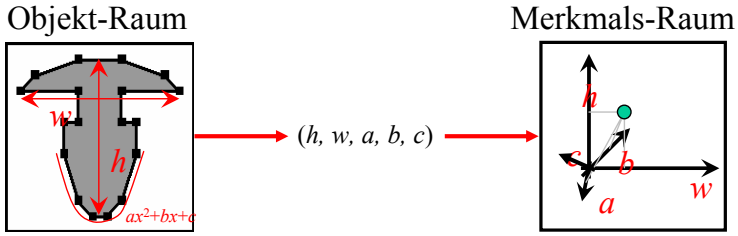
Räume und Distanzfunktionen

- Hauptachsenanalyse eine Menge *DB* von *Euklidischen Vektoren*
  - Kovarianz-Matrix: $\Sigma_{DB} = \frac{1}{|DB|} \sum_{o \in D} (o - \mu_{DB})(o - \mu_{DB})^{\mathbf{T}}$
  - Die Matrix wird zerlegt in
    - eine Orthonormalmatrix V = [e_1,...,e_d] (Eigenvektoren)
    - und eine Diagonalmatrix Λ = diag (λ_1,...,λ_d) (Eigenwerte)
    - so dass gilt: $\Sigma_{DB} = V \Lambda V^{\mathbf{T}}$
  - Interpretation:
    - Eigenvektoren:
      Hauptausrichtung der Datenpunkte in *DB*
    - Eigenwerte:
      Varianz der Datenpunkte in *DB* entlang der entspr. Eigenvektoren

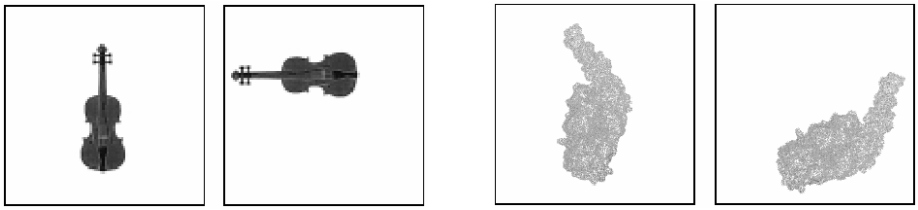


Feature-Transformationen für räumliche Objekte

- Feature Transformation für räumliche Objekte (CAD-Daten, Proteine, …)



  - Invarianzen
    - Gleichheit (oder Ähnlichkeit) von Formen unabhängig von Lage und Orientierung im Raum
    - Beispiele gleicher Formen im 2D und im 3D:

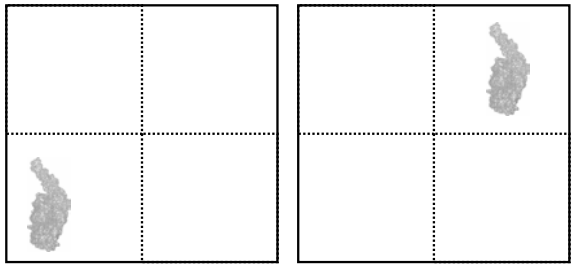


    - Erwünscht:
      - Kanonische Darstellung, d.h. ohne Lage- und Orientierungsinformation
      - Verallgemeinerung auf andere Objekteigenschaften

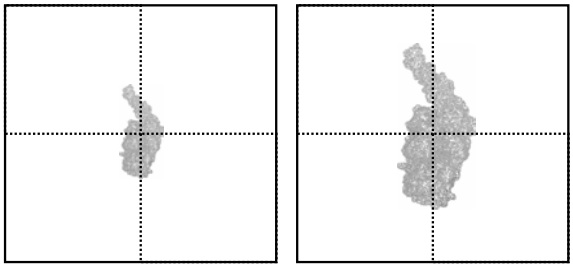
Feature-Transformationen für räumliche Objekte

  - Die wichtigsten Invarianzen

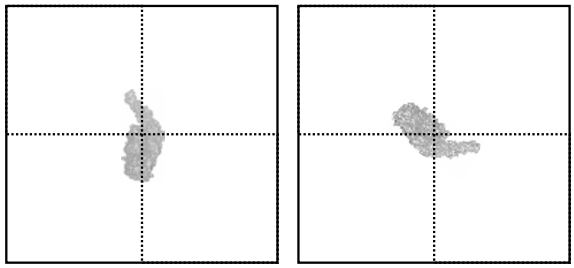
Translation



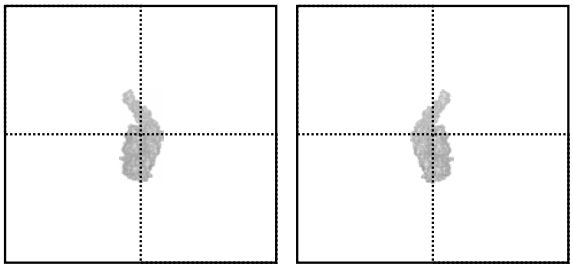
Skalierung



Rotation

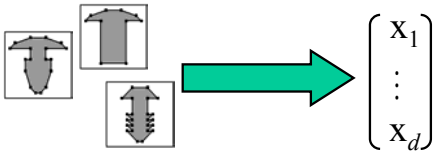


Spiegelung

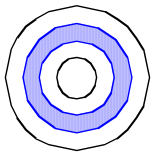


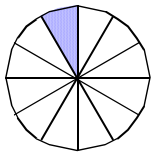
Feature-Transformationen für räumliche Objekte

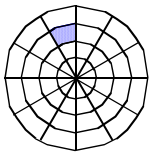
- Volume Model [Ankerst, Kastenmüller, Kriegel, Seidl 99]



  - Applikationen: CAD, Protein 3D-Strukturen
  - Idee: *Formhistogramme* für 3D Objekte
    - Partitioniere den 3D-Raum in Zellen (Histogramm-Bins).
    - Bestimme den Anteil an Punkten des Objektes pro Zelle (normiertes Histogramm).
    - Durch die Normierung werden die Histogramme unabhängig von der Punktedichte.
  - Partitionierungen

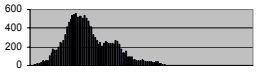

Schalenmodell


Sektorenmodell

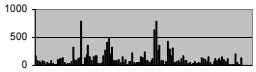

kombiniertes Modell

  - Beispiel

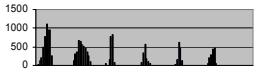

Seryl-tRNA Synthetase



Schalenmodell (120 Schalen)



Sektorenmodell (122 Sektoren)



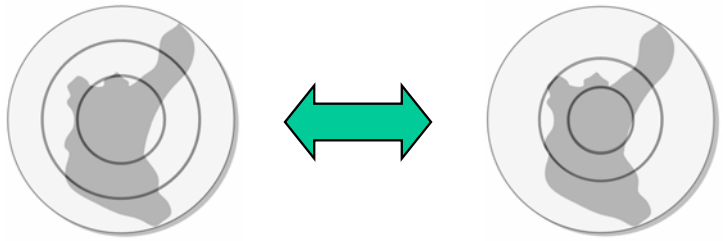
kombiniertes Modell (20 Schalen, 6 Sektoren)

Feature-Transformationen für räumliche Objekte

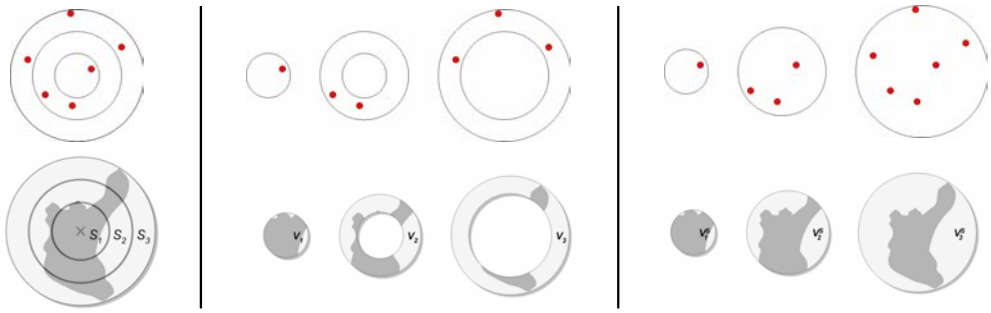
  - Formale Definition der Histogramme
    - *Schalenmodell*: Definiere die Bins über den Abstand zum Mittelpunkt, d.h. Anzahl der Punkte auf der jeweiligen Schale.
    - *Sektorenmodell*: Anzahl der Punkte im jeweiligen Sektor.
    - *Kombiniertes Modell*: Synthese aus Schalen- und Sektorenmodell.
  - Invarianzen
    - Translationsinvarianz durch Lagenormierung: Verschiebung des Schwerpunkts eines Objektes in den Ursprung.
    - Rotationsinvarianz durch Hauptachsentransformation:
      - Drehung der Objekte, so dass die Hauptachsen auf den Koordinatenachsen liegen.
      - unnötig beim Schalenmodell, dieses ist inhärent rotationsinvariant.

Feature-Transformationen für räumliche Objekte

- Verbesserung der Formhistogramme [Aßfalg, Kriegel, Kröger, Pötke 05]
  - Proportionale Aufteilung

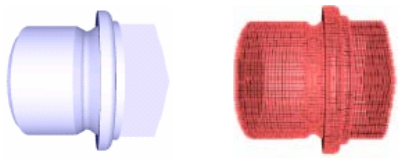


  - Redundante Zuordnung zu den Bins

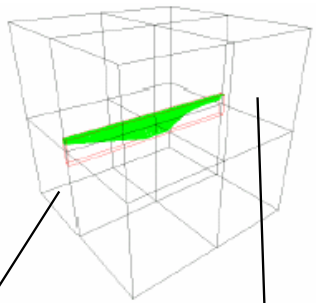


Feature-Transformationen für räumliche Objekte

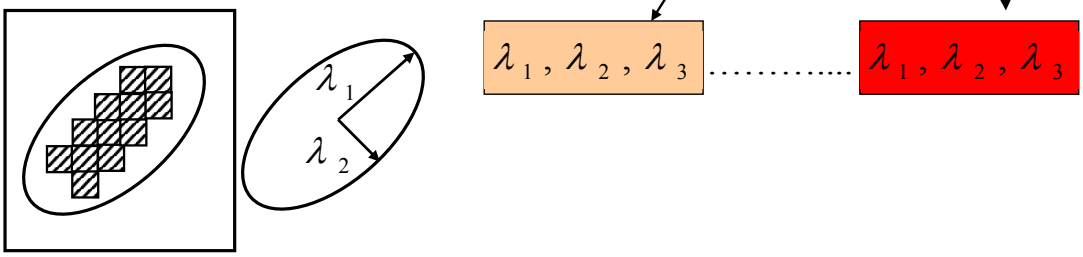
- Eigenvalue Model [Kriegel, Kröger, Mashael, Pfeifle, Pötke , Seidl 03]
  - Volumen-Diskretisierung durch Voxel (3dimensionale Pixel)



  - Würfelförmige Partitionierung der Bounding Box

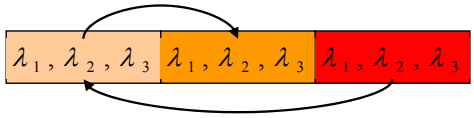


  - Bestimmung der Eigenwerte des Voxelinhaltes jeder Zelle



Feature-Transformationen für räumliche Objekte

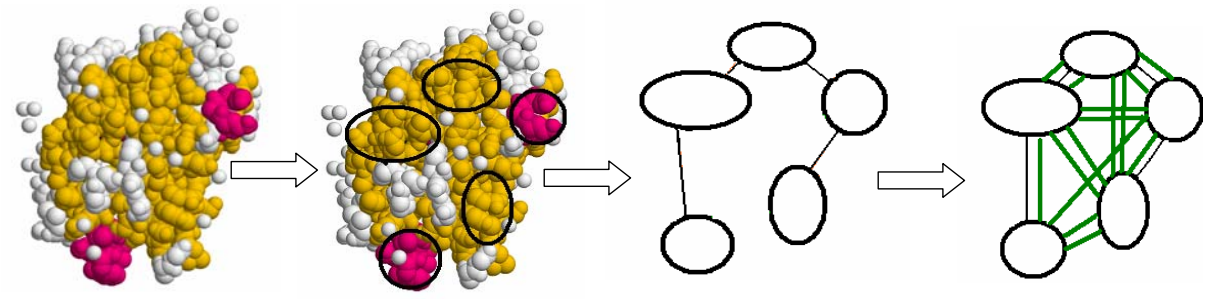
- Invarianzen
  - Translationsinvarianz durch Lagenormierung:
    Verschiebung des Schwerpunkts eines Objektes in den Ursprung.
  - Skalierungsinvarianz durch Voxelisierung der Bounding Box/Bounding Cube des Objekts mit immer gleicher Voxelauflösung
  - Rotationsinvarianz
    - Hauptachsentransformation (völlig rotationsinvariant, aber bei manchen Objekten sensitiv gegenüber kleinen Änderungen)
    - CAD Objekte oft in „vernünftiger“ Lage durch Konstrukteur abgespeichert, dann besser 90-Grad-Rotationsinvarianz: Zur Laufzeit werden die 24 Würfelpositionen durch Permutation der Merkmalsvektor-Elemente simuliert, die Distanz zweier Objekte ist das Minimum über 24 Distanzen



  - Reflektionsinvarianz
    - Betrachte 48 statt 24 Permutationen zur Laufzeit (incl. Spiegelung des Würfels)

Feature-Transformationen für räumliche Objekte

- Protein Datenbanken [Borgwardt, Ong, Schönauer, Vishwanathan, Smola, Kriegel 05]
  - Idee:
    - Graphmodel für Protein 3D-Strukturen
    - Knoten: Untereinheiten von Proteinen (secondary structure elements)
    - Kanten: Nachbarschaft von Untereinheiten innerhalb der 3D-Struktur und entlang der Aminosäure Sequenz.



protein data | secondary structure elements | sequence | structure