

Skript zur Vorlesung
Knowledge Discovery in Databases
im Wintersemester 2010/2011

Kapitel 2: Merkmalsräume

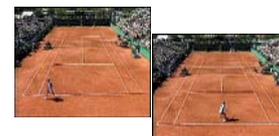
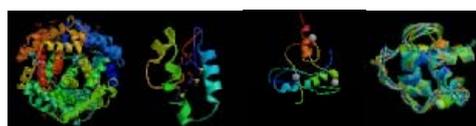
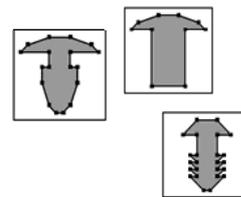
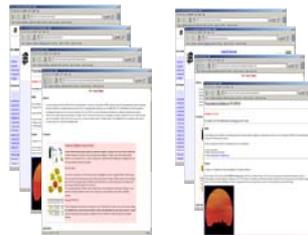
Vorlesung+Übungen:
PD Dr. Peer Kröger, Dr. Arthur Zimek

Skript © 2003 Johannes Aßfalg, Christian Böhm, Karsten Borgwardt, Martin Ester, Eshref Januzaj, Karin Kailing, Peer Kröger, Jörg Sander, Matthias Schubert, Arthur Zimek

[http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_I_\(KDD_I\)](http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_I_(KDD_I))

Merkmalsräume

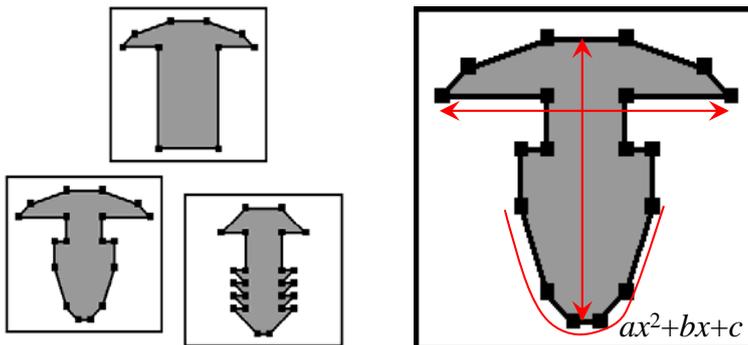
- Motivation:
 - Zentrales Konzept beim Data Mining: Ähnlichkeit von Datenbankobjekten
 - Clustering: Zusammenfassen *ähnlicher* Objekte in Gruppen
 - Klassifikation: Zuordnung von Objekten zu einer Klasse *ähnlicher* Objekte
 - Definition einer geeigneten Distanzfunktion auf Datenbankobjekten nicht 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:

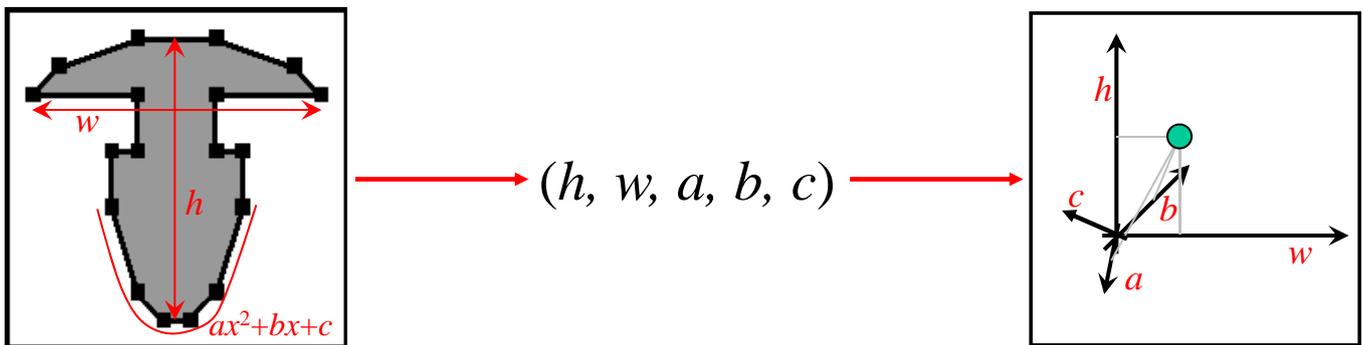


Mögliche Merkmale:

Beispiel: CAD-Zeichnungen (cont.)

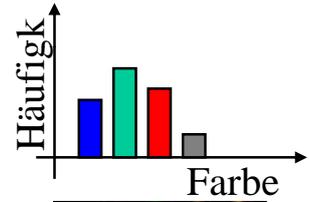
Objekt-Raum

Merkmals-Raum

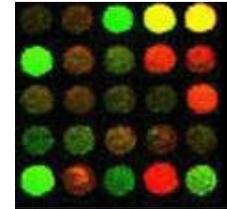


- Im Kontext von statistischen Betrachtungen werden die Merkmale häufig auch als *Variablen* bezeichnet
- Die ausgewählten Merkmale werden zu Merkmals-Vektoren (*Feature Vector*) zusammengefasst
- Der Merkmalsraum ist häufig hochdimensional (im Beispiel 5-dim.)

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

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?)
- Gütekategorie
- Altersklasse

Metrisch

Charakteristik:

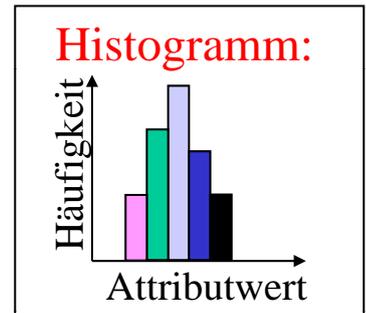
Sowohl Differenzen als auch Verhältnisse zwischen den Werten sind aussagekräftig. Die Werte können diskret oder stetig sein.

Beispiele:

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

Sei x_1, \dots, 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: $f(a) = h(a) / n$



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: $\sigma = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})^2}$

Kontingenztafel

- 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

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

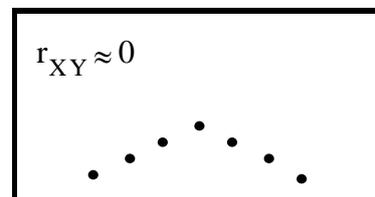
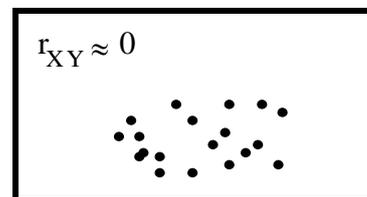
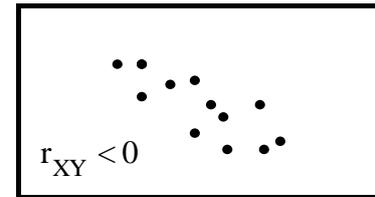
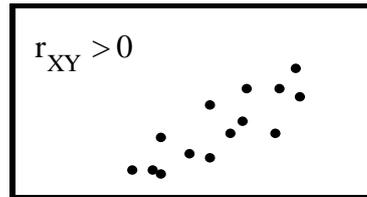
- Wie sollten die relativen Häufigkeiten verteilt sein, wenn die beiden Merkmale keinerlei Abhängigkeit besitzen? $\frac{h_{ik}}{n} = \frac{h_{i.}}{n} \cdot \frac{h_{.k}}{n}$
- χ^2 -Koeffizient
Differenz zwischen dem bei Unabhängigkeit erwarteten und dem tatsächlich beobachteten Wert von h_{ij} (Maß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

Korrelationskoeffizient

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

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

- Beispiele



9

Merkmalsraum (Featureraum)

- Intuitiv: ein Wertebereich/Domain mit Distanzfunktion
- Formal: Featureraum $\mathbf{F} = (Dom, dist)$
- Dom ist eine (geordnete) Menge von Merkmalen (Features)
- $dist: Dom \times Dom \rightarrow \mathbb{R}$ ist eine totale (Distanz)-Funktion mit den folgenden Eigenschaften
 - $\forall p, q \in Dom, p \neq q : dist(p, q) > 0$
 - $\forall o \in Dom : dist(o, o) = 0$
 - $\forall p, q \in Dom : dist(p, q) = dist(q, p)$

Reflexivität

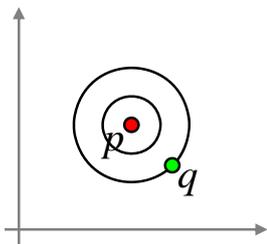
Symmetrie

- Metrischer Raum
 - Formal: Metrischer Raum $\mathbf{M} = (Dom, dist)$ mit den folgenden Eigenschaften
 - \mathbf{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 $\mathbf{E} = (Dom, dist)$ 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“

• Ähnlichkeit von Feature Vektoren (Euklidische Vektoren)

Euklidische Norm (L_2):

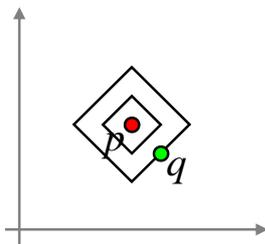
$$dist_1 = ((p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots)^{1/2}$$



Natürlichstes Distanzmaß

Manhattan-Norm (L_1):

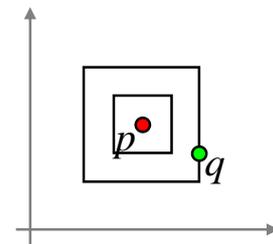
$$dist_2 = |p_1 - q_1| + |p_2 - q_2| + \dots$$



Die Unähnlichkeiten der einzelnen Merkmale werden direkt addiert

Maximums-Norm (L_∞):

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

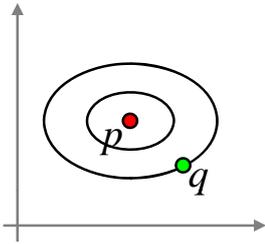


Die Unähnlichkeit des am wenigsten ähnlichen Merkmals zählt

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

Gewichtete Euklidische Norm:

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



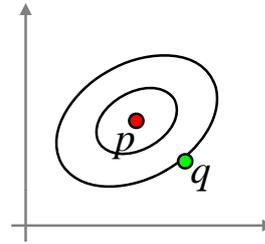
Häufig sind die Wertebereiche der Merkmale deutlich unterschiedlich.

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:

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



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

13

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 durchschnittlichen 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

14

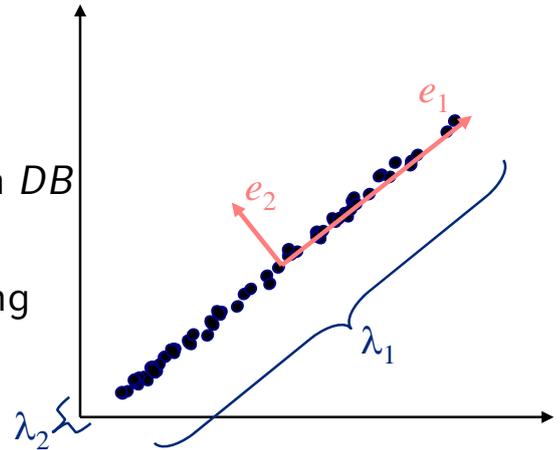
Hauptachsenanalyse eine Menge DB von *Euklidischen Vektoren*

– Kovarianz-Matrix:
$$\Sigma_{DB} = \frac{1}{|DB|} \sum_{o \in DB} (o - \mu_{DB})(o - \mu_{DB})^T$$

- Die Matrix wird zerlegt in
- eine Orthonormalmatrix $V = [e_1, \dots, e_d]$ (Eigenvektoren)
 - und eine Diagonalmatrix $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$ (Eigenwerte)
 - so dass gilt: $\Sigma_{DB} = V \Lambda V^T$

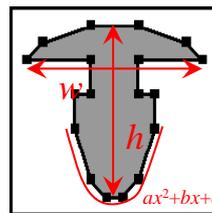
– Interpretation:

- Eigenvektoren:
Hauptausrichtung der Datenpunkte in DB
- Eigenwerte:
Varianz der Datenpunkte in DB entlang der entspr. Eigenvektoren



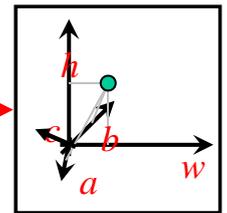
Feature Transformation für räumliche Objekte (CAD-Daten, Proteine, ...)

Objekt-Raum



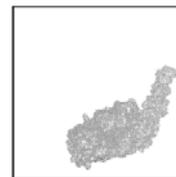
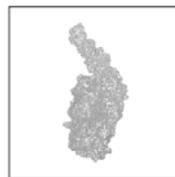
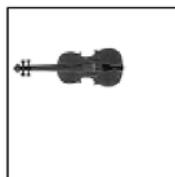
(h, w, a, b, c)

Merkmals-Raum



– Invarianzen

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

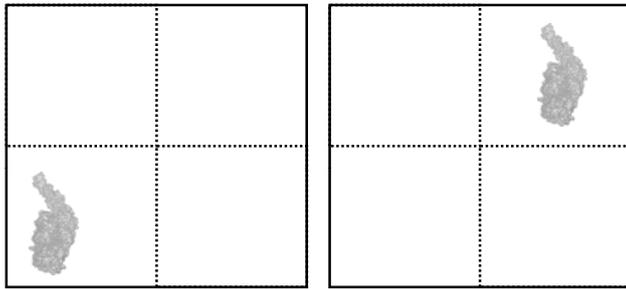


• Erwünscht:

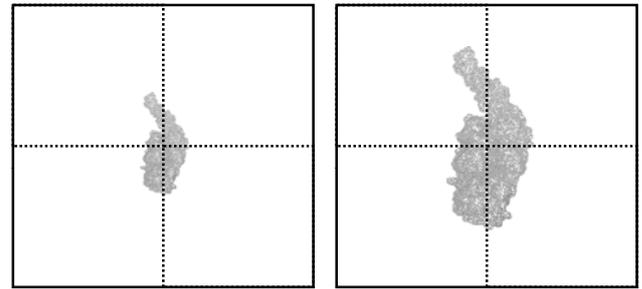
- Kanonische Darstellung, d.h. ohne Lage- und Orientierungsinformation
- Verallgemeinerung auf andere Objekteigenschaften

Die wichtigsten Invarianzen

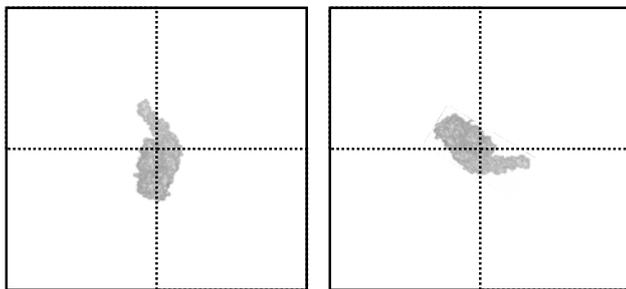
Translation



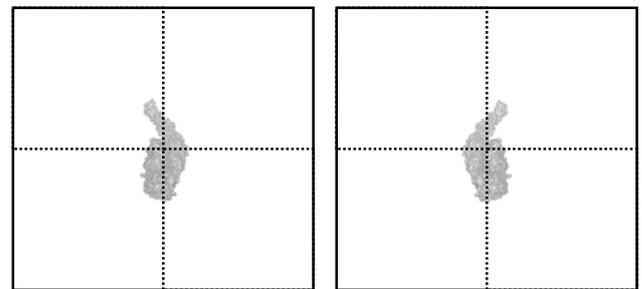
Skalierung



Rotation

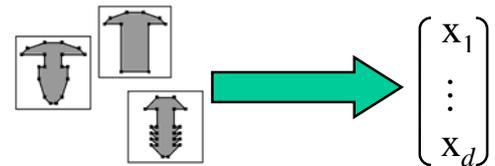


Spiegelung



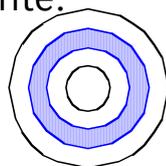
Volume Model [Ankerst, Kastenmüller, Kriegel, Seidl 99]

- Applikationen: CAD, Protein 3D-Strukturen
- Idee: *Formhistogramme* für 3D Objekte

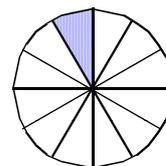


- Partitioniere den 3D-Raum in Zellen (Histogramm-Bins).
- Bestimme den Anteil an Punkten des Objektes pro Zelle (normiertes Histogramm).
- Durch die Normierung werden die Histogramme unabhängig von der Punktedichte.

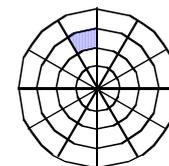
- Partitionierungen



Schalenmodell

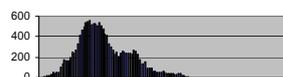


Sektorenmodell



kombiniertes Modell

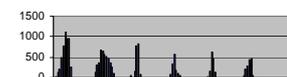
- Beispiel



Schalenmodell (120 Schalen)



Sektorenmodell (122 Sektoren)

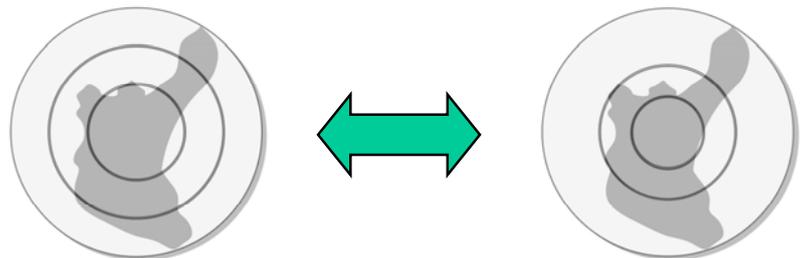


kombiniertes Modell (20 Schalen, 6 Sektoren)

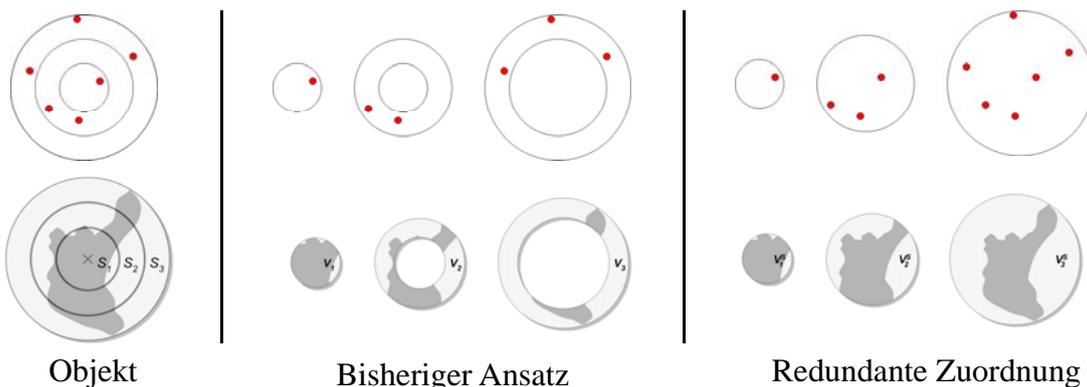
- Formale Definition der Histogramme
 - *Schalenmodell*: Definiere die Bins über den Abstand zum Mittelpunkt, d.h. Anzahl der Punkte auf der jeweiligen Schale.
 - *Sektorenmodell*: Anzahl der Punkte im jeweiligen Sektor.
 - *Kombiniertes Modell*: Synthese aus Schalen- und Sektorenmodell.
- Invarianzen
 - 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.

19

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



- Redundante Zuordnung zu den Bins



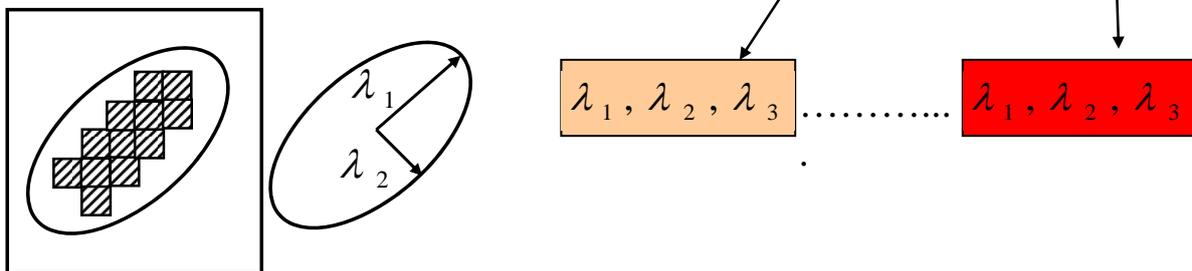
20

Eigenvalue Model [Kriegel, Kröger, Mashael, Pfeifle, Pötke, Seidl 03]

- Volumen-Diskretisierung durch Voxel (3dimensionale Pixel)

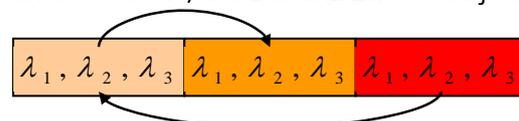


- Würfelförmige Partitionierung der Bounding Box
- Bestimmung der Eigenwerte des Voxelinhaltes jeder Zelle



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



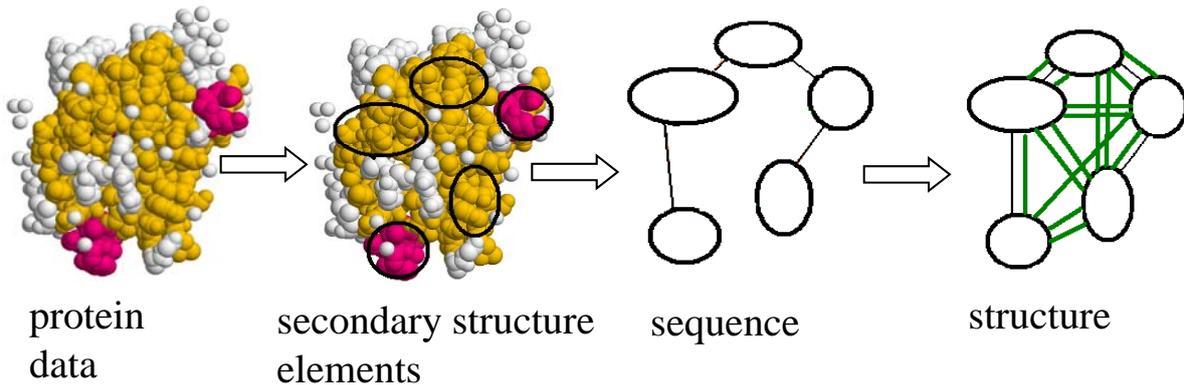
- Reflektionsinvarianz

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

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.

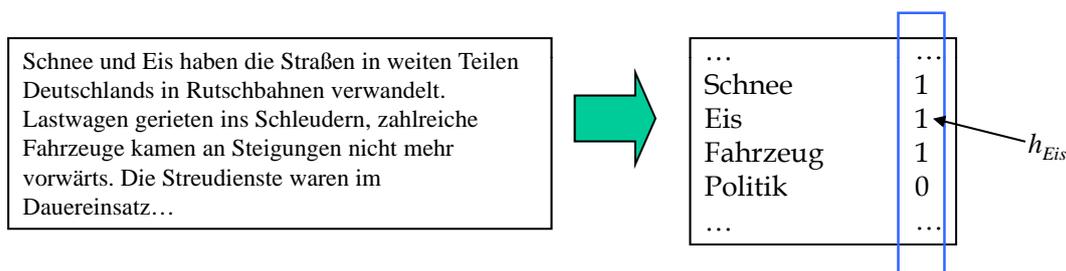


- *Text als Mengen/Vektoren von Termen: („Bag-Of-Words“)*

– Term:

- einzelnes Wort („Schnee“, „Eis“..)
oder
- zusammenhängendes Satzfragment („nicht mehr vorwärts“..)

– Transformation eines Dokuments D in Vektor $r(D) = (h_1, \dots, h_d)$
 $h_i \geq 0$: die Häufigkeit des Terms t_i in D



- Probleme im Textmining
 1. Viele Wörter nutzlos (z.B. er, sie, es, und, als, der, dies, das..)
 2. Wörter haben gleichen Wortstamm („gehen“ „ging“)
 3. Sehr hochdimensionale Featureräume (häufig $d > 10.000$)
 4. Nicht alle Terme sind gleich wertvoll
 5. Die meisten Termhäufigkeiten $h_i = 0$ („sparse feature space“)
- weitere Probleme aus der Linguistik:
 - unterschiedliche Wörter haben gleiche Bedeutung
„laufen“ \Leftrightarrow „rennen“
 - Wörter haben mehrere Bedeutungen
„Maus“: Computermouse, Nagetier...

- Problem 1: Viele Wörter nutzlos (z.B. er, sie, es, und, als, der, dies, das..)
 - Lösung: Streichen solcher Terme (Stopwords)
Für alle Sprachen werden Stopwordlisten im WWW publiziert.
- Problem 2: Wörter haben gleichen Wortstamm („gehen“ „ging“)
 - Lösung: Stemming
Worte auf Wortstamm rückführen (z.B. lief, läuft, lauft => laufen)
Im Englischen algorithmisches Stemming möglich.
(Porters Stemming Algorithms: <http://tartarus.org/~martin/PorterStemmer/index.html>)
In anderen Sprachen werden Dictionaries benötigt, die die Wortstämme zu den Vokabeln enthalten.

- Problem 3: Sehr viele Terme müssen betrachtet werden.
 - Lösung: Auswahl der wichtigsten Features („Feature Selection“)
 - Beispiel: Mittlere Dokumentenhäufigkeit
 - Sehr häufige Terme kommen scheinbar in allen Dokumenten vor
=> Vorkommen unterscheidet kaum Dokumente
 - Sehr seltene Terme kommen nur in Bruchteil der Dokumente vor
=> Nichtvorkommen unterscheidet kaum Dokumente

Vorgehen:

1. Berechne Dokumenthäufigkeit für alle Terme t_i : $DF(t_i) = \frac{|Dok_t_i|}{|ALL_Doks|}$
2. Sortiere Terme nach $DF(t_i)$ und Vergebe Rang $rank(t_i)$
3. Sortiere Terme nach $score(t_i) = DF(t_i) \cdot rank(t_i)$
z.B. $score(t_{23}) = 0.82 \cdot 1 = 0.82$
 $score(t_{17}) = 0.75 \cdot 2 = 1.5$
4. Wähle die k Terme mit dem größten Wert für $score(t_i)$

Rank	Term	DF
1.	t_{23}	0.82
2.	t_{17}	0.65
3.	t_{14}	0.52
4.

27

- Problem 4: Nicht alle Terme sind gleich wertvoll.
 - Idee:
 1. Gewichte seltene Terme höher als häufige.
 2. Gewichte häufig in einem Dokument auftretende Terme höher als solche die nur einmal vorkommen.
 - Lösung: TF-IDF (Term Frequency · Inverse Document Frequency)
Berücksichtige sowohl die relative Anzahl der Vorkommen im Dokument als auch die Seltenheit des Terms.

$$TF(t, d) = \frac{n(t, d)}{\sum_{w \in d} n(w, d)} \quad \text{relative Häufigkeit von } t \text{ in } d$$

$$IDF(t) = \frac{|DB|}{|\{d \mid d \in DB \wedge t \in d\}|} \quad \text{inverse Häufigkeit von } t \text{ bzgl. aller Dokumente}$$

Featurevektor mit TF IDF : $r(d) = (TF(t_1, d) \cdot IDF(t_1), \dots, TF(t_n, d) \cdot IDF(t_n))$

28

- Problem 5: die meisten Termhäufigkeiten $h_i = 0$
 => *Euklidische Abstände sehr ähnlich*
 - Lösung: Verwendung anderer Abstandsmaße
 Idee: Verwende Terme, die beide Dokumente (D_1, D_2) gemeinsam haben.

Jaccard Coefficient: Dokumente als Termmengen

$$d_{Jaccard}(D_1, D_2) = 1 - \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|}$$

Cosinus Coefficient: Abstand für Wortvektoren (evtl. TF IDF)

$$d_{cosinus}(D_1, D_2) = 1 - \frac{\langle D_1, D_2 \rangle}{\|D_1\| \cdot \|D_2\|} = 1 - \frac{\sum_{i=0}^n (d_{1,i} \cdot d_{2,i})}{\sqrt{\sum_{i=0}^n d_{1,i}^2} \cdot \sqrt{\sum_{i=0}^n d_{2,i}^2}}$$