Applied Analytical Data Science

Teil 8: Clustering und Subgruppenentdeckung

Dr. Jörg-Uwe Kietz,
Vorlesung an der Univ. Zürich,

Mittwoch, 14:00-15:45 Uhr Vorlesung,
16:00-17:30 Uhr Übung

http://www.kietz.ch/AADS/

Heutiges Programm

- Clustering
  - Definition, Beispiel und Aufgaben
  - k-means-Clustering
  - Überblick über weitere Ansätze
  - COBWEB: inkrementelles hierarchisches clustering
- Subgruppenentdeckung (Deviation Detection)
  - Definition, Beispiel und Aufgaben
  - Statistisch signifikante Subgruppen
  - Effiziente Suche nach Subgruppen
  - Zusammenhänge und Unterschiede zu Klassifikation, Clustering und Assoziationen

Clustering Definition

Gegeben:

Eine Menge von Instanzen, beschrieben durch

- Eingabe Attribute

Gesucht:

Ein Aufteilung der Instanzen in Klassen (ein Clustering), so dass

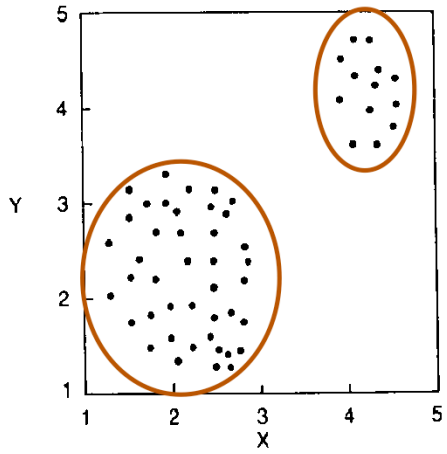
- Instanzen innerhalb einer Klasse möglichst ähnlich sind.
- Instanzen verschiedener Klassen möglichst unähnlich sind.
- eine Beschreibung der Klassen

Aufgaben fürs Clustering

Marketing

- Kunden Segmentierungen, z.B.
  - Zielgruppen
  - Gruppen mit typischem Verhalten
- Daten Überblick/Kompression (k Klassen statt n Instanzen n >> k)
- Entdeckung von
  - Einflussreichsten Attributen
  - Abhängigkeiten zwischen Attributen
  - Datenfehlern/Ausreissern

Clustering Beispiel



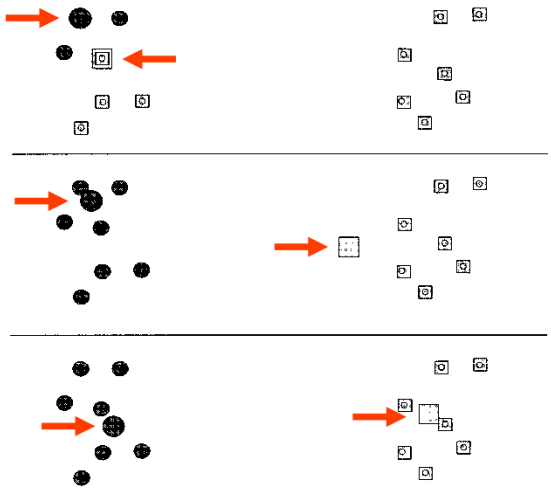
- Ein einfaches Verfahren um solche Cluster zu finden ist **k-means**.
- **k-means** arbeitet auf skalaren Daten.
- k, die Anzahl der Cluster, wird vorgegeben.
- **k-means**-Clustering erzeugt k Cluster indem es k Zentrums-punkte findet.

k-means Algorithmus

proc K-MEANS(S, k)
  $Z = \{z_1, \ldots, z_k\} := k$ zufällig gewählte Punkte aus S
  while Qualität wird besser
    $C_i := \{s \in S | i = argmin_{i=1,\ldots,k} dist(s, z_i)\}$ für $i = 1, \ldots, k$
    $z_i := z(C_i)$ für $i = 1, \ldots, k$
    end
  return $\{C_1, \ldots, C_k\}$

$$z(C) := \frac{1}{|C|} \sum_{\mathbf{c} \in C} \mathbf{c}$$

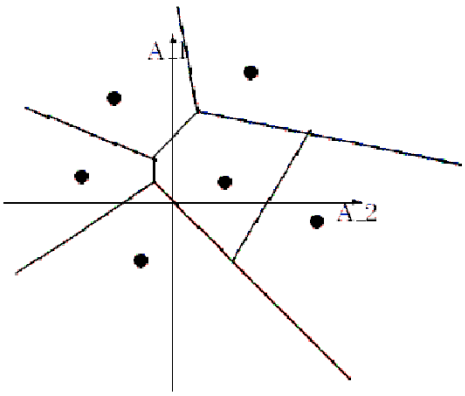
k-means Beispiel



k-means-Clustering

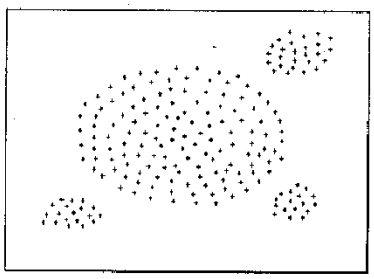
- k-means-Clustering erzeugt k Zentrumspunkte
- die Zuordnung der Instanzen zu den Klassen (die Klassifikation) erfolgt (wie bei der Methode der nächsten Nachbarn) durch die Bestimmung des ähnlichsten Zentrumspunktes.

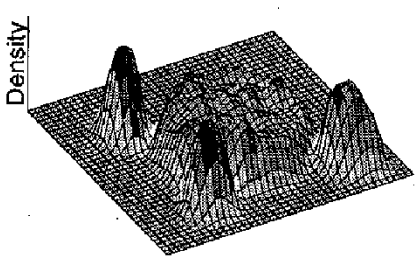
Voronoi-Diagramm für Nachbarschaftsbeziehungen



Density-based Clustering

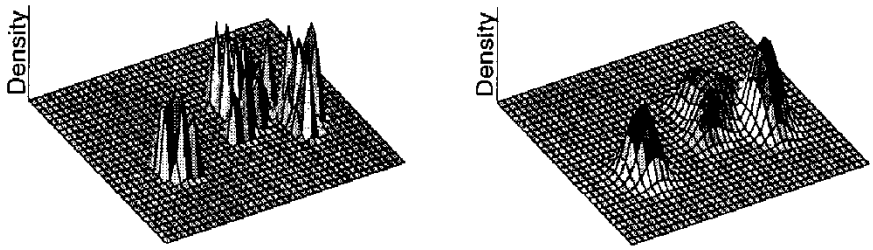
- Eine Dichtematrix wird berechnet, indem für jeden möglichen Punkt (die Auflösung bestimmt den Aufwand!) gezählt wird, wieviele Punkte in der Umgebung liegen.





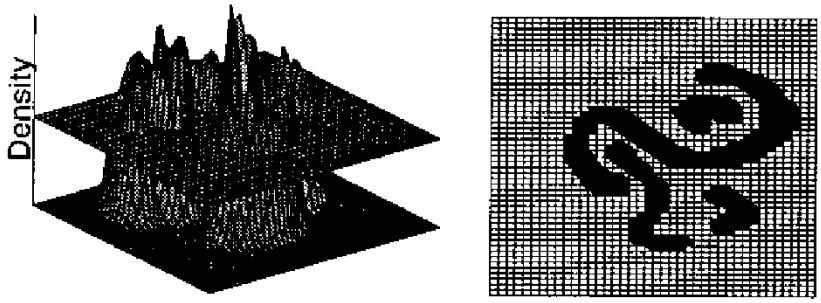
Density-based Clustering

- Die Grösse der Region für die die Dichte bestimmt wird ist entscheident:
  - Ist sie zu klein, „zerfällt“ das Gesamtbild
  - Ist sie zu gross, „verwischt“ das Gesamtbild



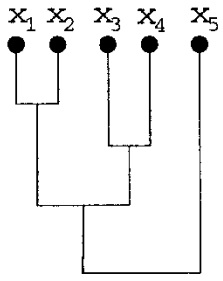
Density-based Clustering

- Nach der Berechnung der Dichtematrix kann die Auflösung, und damit die Anzahl der Cluster, einfach durch „Verschieben“ eines Dichte-Schwellwertes gesteuert werden.



Hierarchiches Clustering

$\{(x_1), (x_2), (x_3), (x_4), (x_5)\}$
$\{(x_1, x_2), (x_3), (x_4), (x_5)\}$
$\{(x_1, x_2), (x_3, x_4), (x_5)\}$
$\{(x_1, x_2, x_3, x_4), (x_5)\}$
$\{(x_1, x_2, x_3, x_4, x_5)\}$

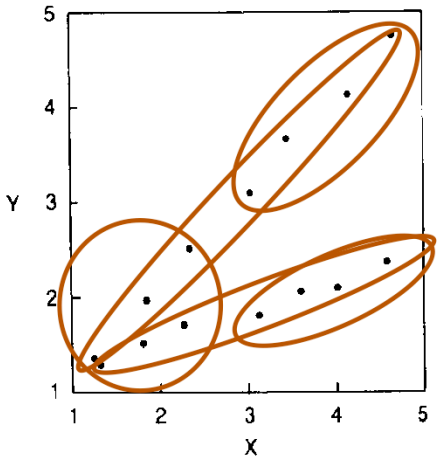


Disjoint

Conjoint

- Hierarchisches Bottom-Up Clustering startet mit den Instanzen und fasst auf jeder Ebene die beiden jeweils ähnlichsten Instanzen/ Cluster zusammen.

Begriffliches Clustering



- Abstandsbasiertes Clustering

  versus

- Begriffliches Clustering

Inkrementelles hierarchiches Clustering

- COBWEB Clustering von nominalen Daten (a)
- CLASSIT Clustering von skalaren Daten (b)
- COBWEB & CLASSIT Integration für gemischte Daten
- Beispiele:

(a)

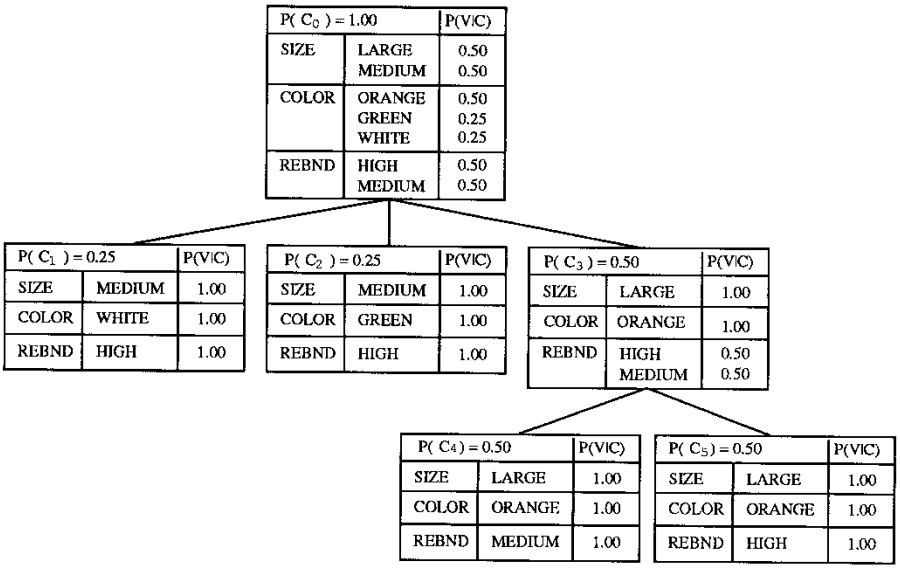
DIAMETER	medium
COLOR	green
REBOUND	high

(b)

DIAMETER	2.52
COLOR	3
REBOUND	0.71

- Beispiele werden inkrementell in eine dabei entstehende Hierarchie von Begriffen einsortiert.

COBWEB: Hierarchie von nominalen Daten



$P(C_0) = 1.00$		P(V|C)
SIZE	LARGE	0.50
	MEDIUM	0.50
COLOR	ORANGE	0.50
	GREEN	0.25
	WHITE	0.25
REBND	HIGH	0.50
	MEDIUM	0.50

$P(C_1) = 0.25$		P(V|C)
SIZE	MEDIUM	1.00
COLOR	WHITE	1.00
REBND	HIGH	1.00

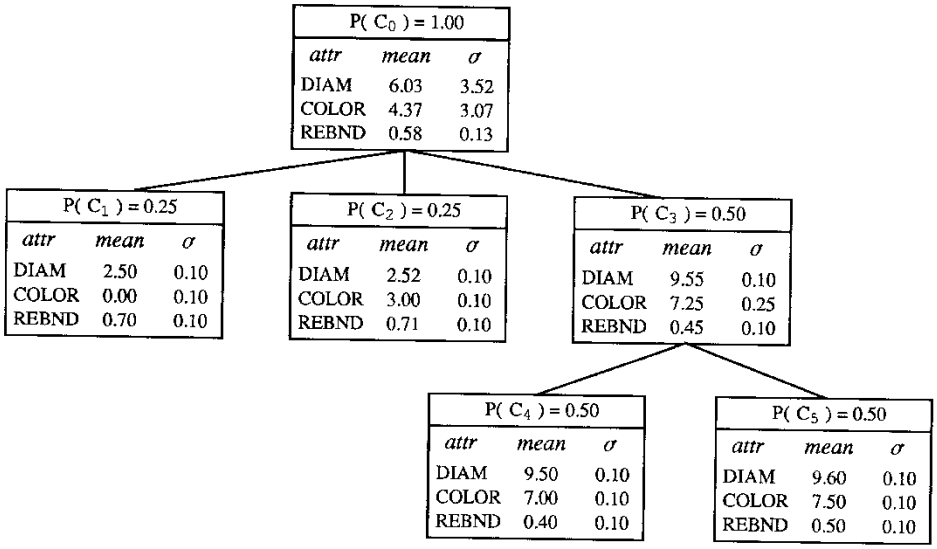
$P(C_2) = 0.25$		P(V|C)
SIZE	MEDIUM	1.00
COLOR	GREEN	1.00
REBND	HIGH	1.00

$P(C_3) = 0.50$		P(V|C)
SIZE	LARGE	1.00
COLOR	ORANGE	1.00
REBND	HIGH	0.50
	MEDIUM	0.50

$P(C_4) = 0.50$		P(V|C)
SIZE	LARGE	1.00
COLOR	ORANGE	1.00
REBND	MEDIUM	1.00

$P(C_5) = 0.50$		P(V|C)
SIZE	LARGE	1.00
COLOR	ORANGE	1.00
REBND	HIGH	1.00

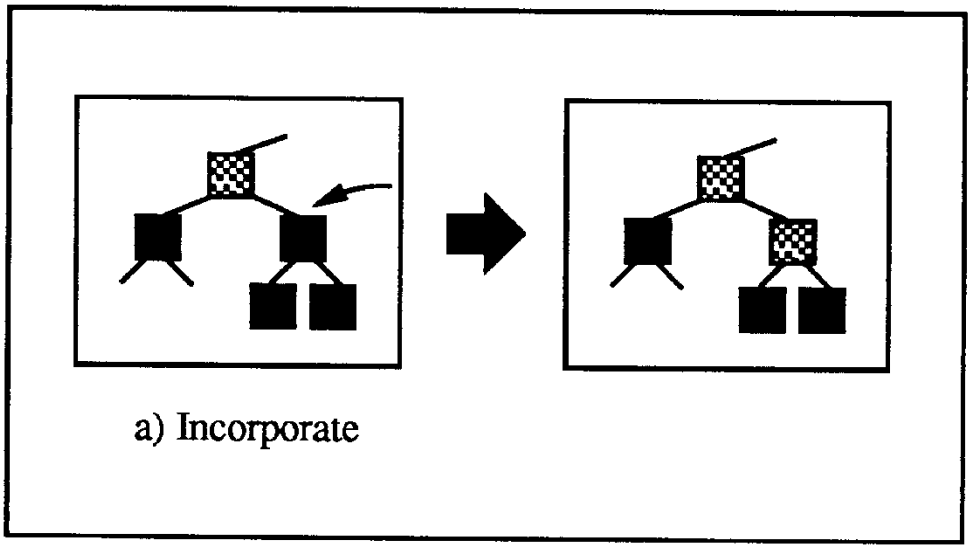
CLASSIT : Hierarchie von skalaren Daten



Inkrementeller Aufbau der Hierarchie

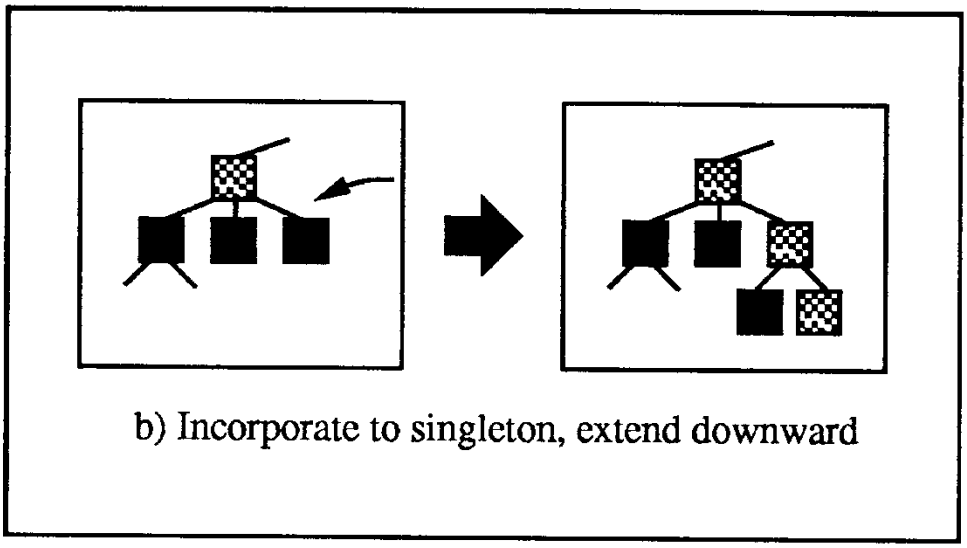
- Instanzen werden eine nach der anderen in die Hierarchie integriert.
- Die Integration startet an der Wurzel und auf jeder Ebene wird die beste Form der Integration ausgewählt, bis die neue Instanz schliesslich als neues Blatt in die Hierarchie einsortiert ist.
- Für die Integration gibt es 5 Operationen:
  - Integration in eine Klasse (2 Operationen: Blatt und nicht Blatt)
  - Einhängen der Instanz als neues Blatt
  - Einhängen in die Vereinigung der beiden besten Möglichkeiten
  - Auflösen eines Knoten und Integration in seinen Unterbaum
- Die beste Form der Integration ist die mit der höchsten Vorhersagekraft.

Inkrementeller Aufbau der Hierarchie



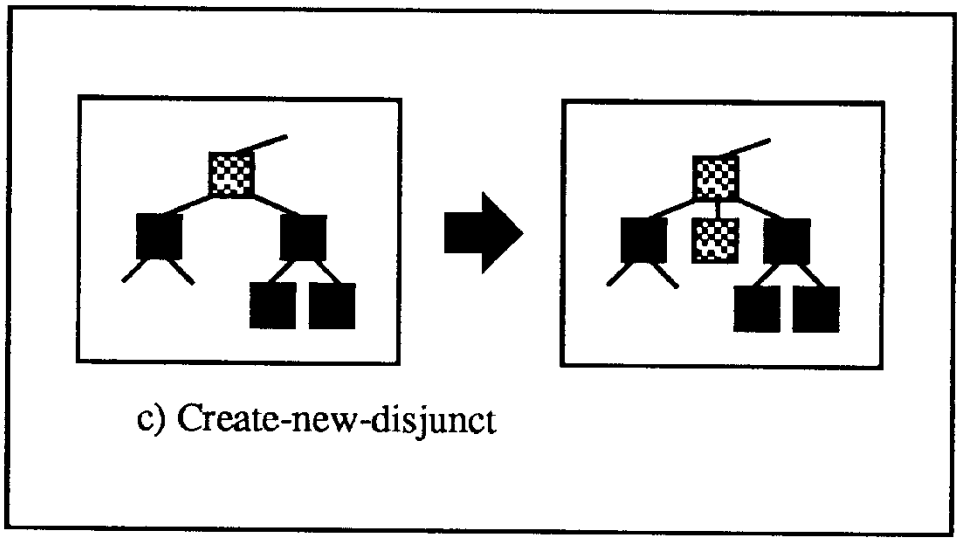
a) Incorporate

Inkrementeller Aufbau der Hierarchie



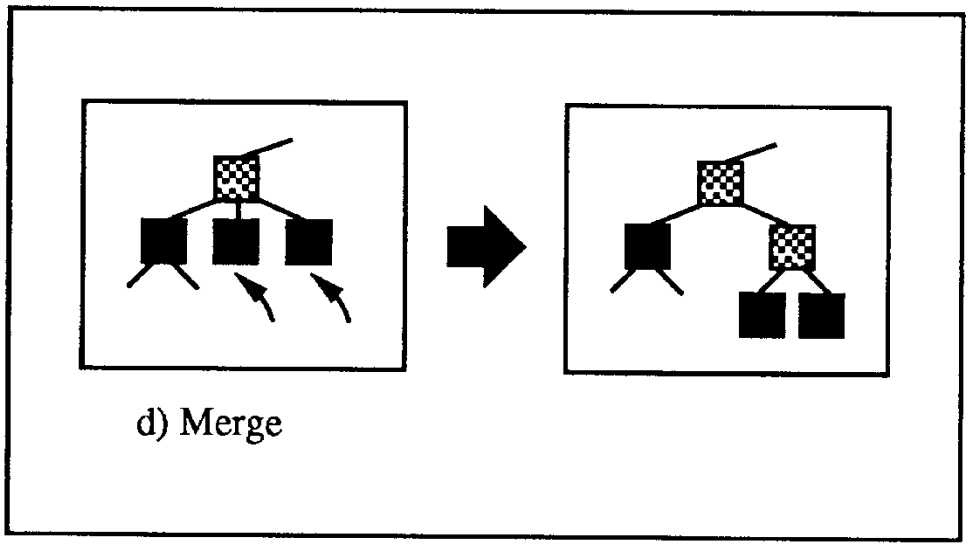
b) Incorporate to singleton, extend downward

Inkrementeller Aufbau der Hierarchie



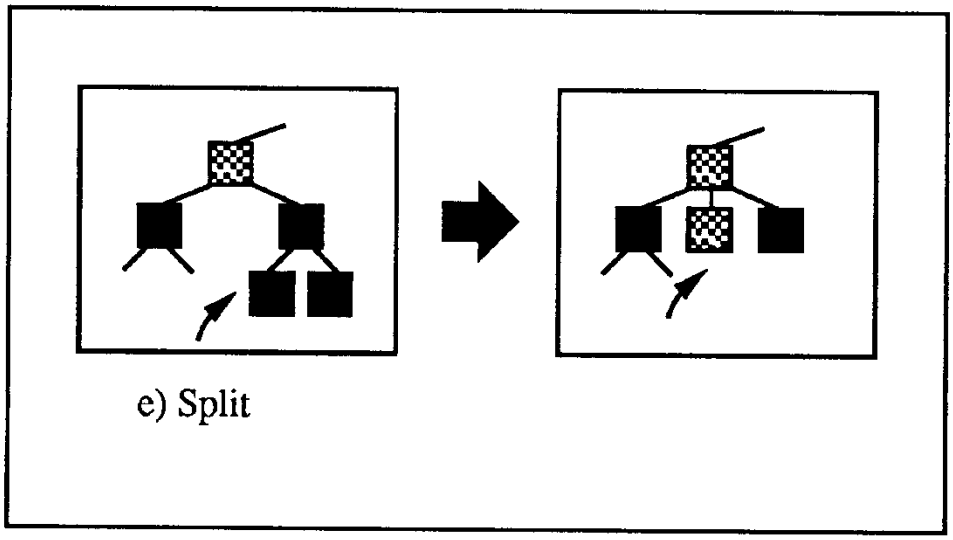
c) Create-new-disjunct

Inkrementeller Aufbau der Hierarchie



d) Merge

Inkrementeller Aufbau der Hierarchie



e) Split

Die Bewertungsfunktion

- Alle durch die verschiedenen Operatoren erzeugten Partitionen (Anzahl der bisherigen Klassen + 3) werden durch die Bewertungsfunktion beurteilt, die mit der besten Bewertung wird ausgewählt.

- Die Bewertungsfunktion ist: $\frac{X - Y}{K}$, mit:

- X: die erwartete Anzahl der korrekt vorherzusagenden Attribute, basierend auf der Klassifikation in K Klassen,
- Y: die erwartete Anzahl der korrekt vorherzusagenden Attribute ohne eine Klassifikation

Die Bewertungsfunktion von COBWEB

- Für die Klassen C_k, dieAttribute A_i mit Werten V_{ij}, ergibt sich erwartete Vorhersage aus den Einzelwahrscheinlichkeiten P
- Y = $\sum_{i=1}^{I}\sum_{j=1}^{J} P(A_i = V_{ij})^2$,
- X = $\sum_{k=1}^{K} P(C_k)\sum_{i=1}^{I}\sum_{j=1}^{J} P(A_i = V_{ij}|C_k)^2$

Und damit für die Gesamtformel

$$\frac{\sum_{k=1}^{K} P(C_k)\sum_i\sum_j P(A_i = V_{ij}|C_k)^2 - \sum_i\sum_j P(A_i = V_{ij})^2}{K}$$

Die Bewertungsfunktion von CLASSIT

- Für numerische Daten wird die Wahrscheinlichkeit von korrekten Vorhersagen mit Hilfe der Normalverteilungsfunktion errechnet:

$$\sum_{j}^{values} P(A_i = V_{ij})^2 \quad \Leftrightarrow \quad \int \frac{1}{\sigma 2\pi} e^{-(\frac{x-\mu}{\sigma})^2} \, dx \quad = \quad \frac{1}{\sigma} \frac{1}{4\sqrt{\pi}}$$

- Die Wahrscheinlichkeit entspricht der Fläche unter der Kurve. Sie ist von der Standardabweichung σ abhängig:

$$\frac{\sum_{k}^{K} P(C_k) \sum_{i}^{I} 1/\sigma_{ik} \; - \; \sum_{i}^{I} 1/\sigma_{ip}}{4K\sqrt{\pi}}$$

- inimale Auflösung A benutzt.

Bewertungsfunktionen für fehlende Werte

- Die Bewertungsfunktion wird normalisiert mit **I** der Anzahl der Attribute mit Werten bei der neuen Instanz.

COBWEB:
$$\frac{\sum_{k=1}^{K} P(C_k)\frac{\sum_i^I \sum_j^J P(A_i=V_{ij}|C_k)^2}{I} - \frac{\sum_i^I \sum_j^J P(A_i=V_{ij})^2}{I}}{K}$$

CLASSIT:
$$\frac{\sum_{k=1}^{K} P(C_k)\frac{\sum_i^I 1/\sigma_{ik}}{I} - \frac{\sum_i^I 1/\sigma_{ip}}{I}}{4K\sqrt{\pi}}$$

- Für gemischte (nominale & skalare) Daten können die Summen auch integriert werden

Eigenschaften von COBWEB/CLASSIT

- COBWEB/CLASSIT brauchen nur einen Table-Scan.
- COBWEB/CLASSIT können Daten inkrementell verarbeiten
- Bei **n** Instanzen in der Datenbank umfasst die Hierarchie ca **2*n** Knoten.

  => Zum Data Mining (grosses **n**) sollten zusätzlich Pruning-Mechanismen eingebaut werden.
- Die Reihenfolge der Datensätze hat einen starken Einfluss auf das Ergebnis. Am besten funktioniert eine echte Zufallsreihenfolge.
- Die Skalierung der skalaren Attribute beeinflusst das Ergebnis stark. Dies ist insbesondere bei gemischten Daten problematisch.

Heutiges Programm

- Clustering
  - Definition, Beispiel und Aufgaben
  - k-means-Clustering
  - Überblick über weitere Ansätze
  - COBWEB: inkrementelles hierarchisches Clustering
- Subgruppenentdeckung (Deviation Detection)
  - Definition, Beispiel und Aufgaben
  - Statistisch signifikante Subgruppen
  - Effiziente Suche nach Subgruppen
  - Zusammenhänge und Unterschiede zu Klassifikation, Clustering und Assoziationen

Definition Subgruppenentdeckung

Gegeben:

Eine Menge von Instanzen, beschrieben durch

- Nominale Eingabeattribute, und
- Ein nominales Zielattribut
- Minimaler Gruppengrösse G_{min} und Anzahl der Subgruppen k

Gesucht:

Die k Subgruppen,

- die mindestens G_{min} Instanzen enthalten, und
- bei denen sich die Verteilung des Zielattributes am signifikantesten von der der Gesamtpopulation unterscheidet.

Subgruppenentdeckung Beispiel

3.5% (1 out of 30)
antworteten auf ein Mailing

ID	Sex	Status	Region	Response
1	male	single	city	no
2	female	family	city	no
3	male	single	city	no
4	male	family	rural	no
5	female	single	city	no
6	male	single	city	no
...				
25	female	single	city	yes
...				
48	male	family	city	yes
...				

Welche Untergruppen haben ein besseres Antwortverhalten?

Positive Abweichungen:

Sex=female, Status=single [9.31%]

Status=single, Region=rural [9.86%]

Region=rural [5.56%]

...

Negative Abweichungen:

Sex=female, Region=city [1.92%]

Status=family [1.91%]

...

Aufgaben für die Subgruppenentdeckung

Marketing

- Verbesserung von Mailing-Aktionen
- Potentielle neue Kunden
- Produkt-Zielgruppen

Geschäftskennzahlen

- Die grössten Veränderungen

Produktentwicklung

- Versicherungen: Gruppen mit signifikant verschiedenem Risiko.

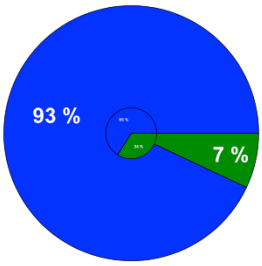
⇒ Immer wenn die Vorhersage einer Klasse verbessert werden soll, aber ein echte Klassifikation nicht erreichbar ist.

Signifikanteste Abweichungen der Verteilung

- Die Qualität einer Hypothese q(h) wird berechnet mit $q(h) := \sqrt{g} \cdot |\hat{p} - \hat{p}'|$.
- Wobei g die relative Grösse der Subgruppe von h zur Gesamtpopulation S ist (g = |h| / |S|)
- $\hat{p}$ ist die Wahrscheinlichkeit mit der das Zielattribut in der Subgruppe von h einen bestimmten Wert hat.
- $\hat{p}'$ ist die Wahrscheinlichkeit mit der das Zielattribut in der Gesamtpopulation diesen Wert hat.
- Eine Herleitung der Qualitätsfunktion basierend auf Statistischen Signifikanz Tests findet sich in [Morik/Wrobel/Joachims 2000].

Signifikanteste Subgruppen

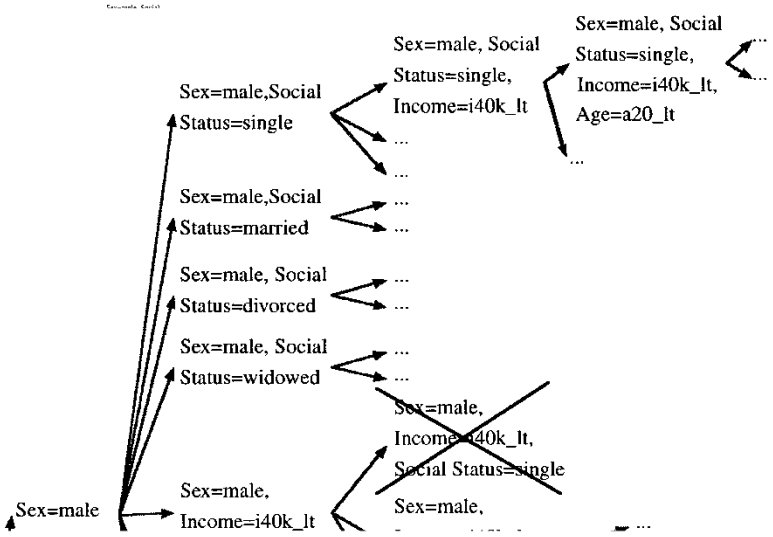
- Die Gesamtpopulation hat eine Verteilung des Zielattributes (beantwortete Werbesendung) von z.B. 93% : 7 %
- Nullhypothese: Ein zufällige Stichprobe hat die gleiche Verteilung.
- Je grösser die Stichprobe und je grösser die Abweichung der Verteilung, desto signifikanter unterscheidet sich die Subgruppe von der Gesamtpopulation, z.B.:
- GUTER_KUNDE=YES Verteilung 34.1%: 65.9%
  Subgruppengrösse: 3948 Haushalte (von 151556)



Suchraum der Subgruppenentdeckung

- Gruppen entsprechen allen möglichen Konjunktionen von Attribut-Wert Paaren.
- Ausgehend von der leeren Konjunktion (ganze Population) werden die Gruppen durch Hinzunehmen von Bedingungen verfeinert.
  - Wie bei Apriori gilt: Wenn die Grösse g der Gruppe einer Konjunktion kleiner ist als G_{min}, hat auch keine der Subgruppen genügende Grösse.
  - Die Qualität einer Verfeinerung kann nicht besser werden als $q_{max}(h) := \sqrt{g} \cdot \max(\hat{p}', 1 - \hat{p}')$, da g nur kleiner werden kann.
- Beides wird zum Vermeiden unnützer Suche benutzt.

Suchraum der Subgruppenentdeckung



Subgruppenentdeckung Vergleich

Klassifikation :

- Die durch Subgruppenentdeckung gefundenen Regeln sind nicht korrekt zur Vorhersage des Zielattributes, und nicht alle Instanzen werden abgedeckt.
- Subgruppenentdeckung kann auch benutzt werden, wenn die vorhandenen Attribute nicht aussagekräftig genug sind eine Klassifikation zu bilden.
- Subgruppenentdeckung kann auch benutzt werden, wenn das Zielattribute für eine Klassifikation zu ungleich verteilt ist. (Ein fehlertoleranter Klassifikationslerner wird bei einer 95:5 Verteilung die 5% zu Fehlern erklären.)

Subgruppenentdeckung Vergleich

Assoziationen :

- Assoziationsregeln (über nominalen Attributen) und Subgruppenentdeckung haben Ähnlichkeiten und Unterschiede:
- Das Format der Regeln ist ähnlich.
- Subgruppenentdeckungs Regeln sind auf ein Zielattribut fokusiert, Assoziationsregeln haben alle möglichen Zielattribute
- Support ~ Gruppengrösse.
- Confidence und Qualitätsfunktion sind unterschiedlich und erlauben unterschiedliche Optimierungen der Suche.

Subgruppenentdeckung Vergleich

Clustering:

- Clustering und Subgruppenentdeckung bilden Gruppen der Gesamtpopulation
- Die meisten Clustering-Verfahren bilden Partitionen der Daten, d.h. die einzelnen Gruppen sind disjunkt, und alle Daten sind durch die Partitionen abgedeckt.
- Die durch Subgruppenentdeckung gefundenen Gruppen sind nicht disjunkt (alle Gruppen könnten die gleichen Elemente enthalten) und decken auch nicht die gesamte Population ab.
- Subgruppenentdeckung beschreibt die Gruppen, das tun die meisten Clustering-Verfahren nicht.

Literatur

Data Mining Methoden zur Untergruppenentdeckung & Clustering:

- K. Morik; S. Wrobel; T. Joachims: "Maschinelles Lernen und Data Mining" Beitrag zum »Handbuch KI«, G. Görz, J. Schneeberger und C.-R. Rollinger (Hrsg.), Oldenbourg Verlag, im erscheinen.
  PDF download: http://www-ai.informatik.uni-dortmund.de/LEHRE/VORLESUNGEN/MLRN/SKRIPT/handbuch_ki-ml.pdf

- Witten, I.; Frank, E.: Practical Machine Learning Tools and Techniques with Java implementations, Morgan Kaufmann, 2000.

Literatur

Vertiefende Literatur zu Clustering:

- Gennari, J.; Langley, P.; Fisher, D.: Models of Incremental Concept Formation, Artificial Intelligence Journal, Vol 40, pp 11-61, 1989.
- Cheeseman, P.; Stutz, J.: Bayesian Classification (AutoClass): Theory and Results, In: Fayyad, U.; Piatetetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, 1996.

Literatur

Vertiefende Literatur zu Subgruppenentdeckung:

- Wrobel, S.: An algorithm for multi-relational discovery of subgroups, In: Komorowski, J.; Zytkow, J.: Proc. of the first PKDD-97, Springer Verlag, 1997.
- Klösgen, W.: Explora: A Multipattern and Multistrategy Discovery Assistent, In: Fayyad, U.; Piatetetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, 1996.