Data Mining in speziellen Daten und Data Mining Anwendungen

Vortrag im Rahmen des Seminars
Business Intelligence -Teil II:
Data Mining & Knowledge Discovery

Holger Klus
30.01.2004

Gliederung

- Text Mining
- Image Mining
- Video Mining
- Mining in räumlichen Daten
- Soziale Aspekte
- Zusammenfassung

Text Mining

- Text-Mining-Anwendungen
  - Klassifizieren von Textdokumenten
  - Erkennen von Trends
  - Generieren von Textzusammenfassungen
  - Halbautomatisches Beantworten von Kundenanfragen

Text Mining

- Klassifizieren von Textdokumenten
  - Preprocessing
    - Entfernen von Formatierungszeichen, HTML-Tags oder Ähnliches
    - Entfernen von Stoppworten
      - Artikel, Präpositionen, ...
    - Stammbildung
      - Zusammenfassung von Worten, die in unterschiedlichen syntaktischen Formen im Dokument auftreten, aber vom selben Wort abstammen(Suche, suchen, gesucht, ...)

Text Mining

- Indexierung
  - Auswahl eines Modells zur Repräsentation von Textdokumenten
    - Vektorraummodell
      - Dokumente sind Vektoren von Wörtern
      - Sammlung von Dokumenten wird durch eine Dokumentmatrix repräsentiert, mit

$$\boldsymbol{A} = (a_{ik})$$

Text Mining

- Gewichtung der Worte nach zwei Regeln
  - Je öfter ein Wort im Dokument enthalten ist, desto höher sein Gewicht
  - Je öfter ein Wort in allen Dokumenten enthalten ist, desto geringer sein Gewicht
- Berechnung des Gewichtes durch

$$a_{ik} = f_{ik} * \log\left(\frac{N}{n_i}\right)$$

	D_1	D_2	D_3	D_4
W_1	3,43	0	6,23	0
W_2	1,76	11,76	9,34	0
W_3	0	9,23	0	2,98
W_4	6,99	0	0.92	1,98
W_5	3,78	2,14	0	6,93

Text Mining

- Dimensionsreduktion
  - Problem
    - Sehr große Dokumentmatrix
    - Dokumentmatrix nur dünn besetzt
  - Document Frequency Thresholding
    - Document Frequency: Anzahl Dokumente, in denen ein Wort vorkommt

Text Mining

- Anwendung eines Klassifizierungsalgorithmus
  - kNN (k-Nearest Neighbour)
    - Eingabe: Dokumentvektor ***d***, Trainingsmenge ***D***
      (1) Bestimmen der k ähnlichsten Nachbarn von ***d***
      (2) Gewichtung der Klassen, in denen die k ähnlichsten Nachbarn von ***d*** enthalten sind
    - Berechnung der Ähnlichkeit
      - Euklidischer Abstand

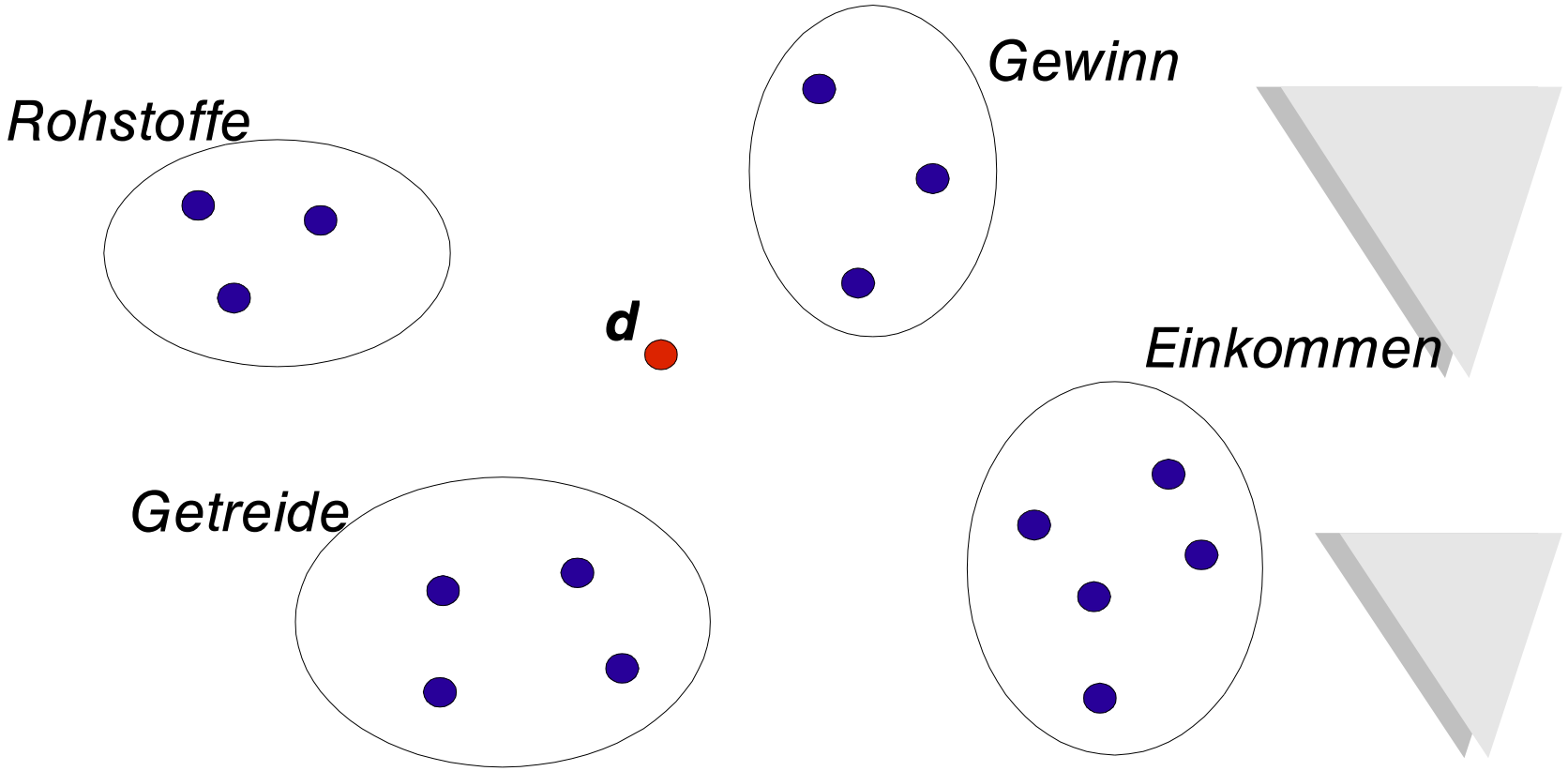
$$\alpha = \arccos\left(\frac{\langle a, b \rangle}{(\|a\| * \|b\|)}\right)$$

Text Mining

- Klassifizieren von Nachrichten in der Reuters-Textsammlung
  - Enthält über 12.000 Nachrichtenartikel
  - Bis zu 135 Kategorien
    - Einkommen, Gewinn, Rohstoffe, ...
- Aufgabe
  - Automatisches Zuordnen von Dokumenten zu Kategorien anhand einer Trainingsmenge
    - 7.000 Trainingsdokumente
    - 2.600 Testdokumente
- Ziel
  - Messung der Effektivität des vorgestellten Verfahrens

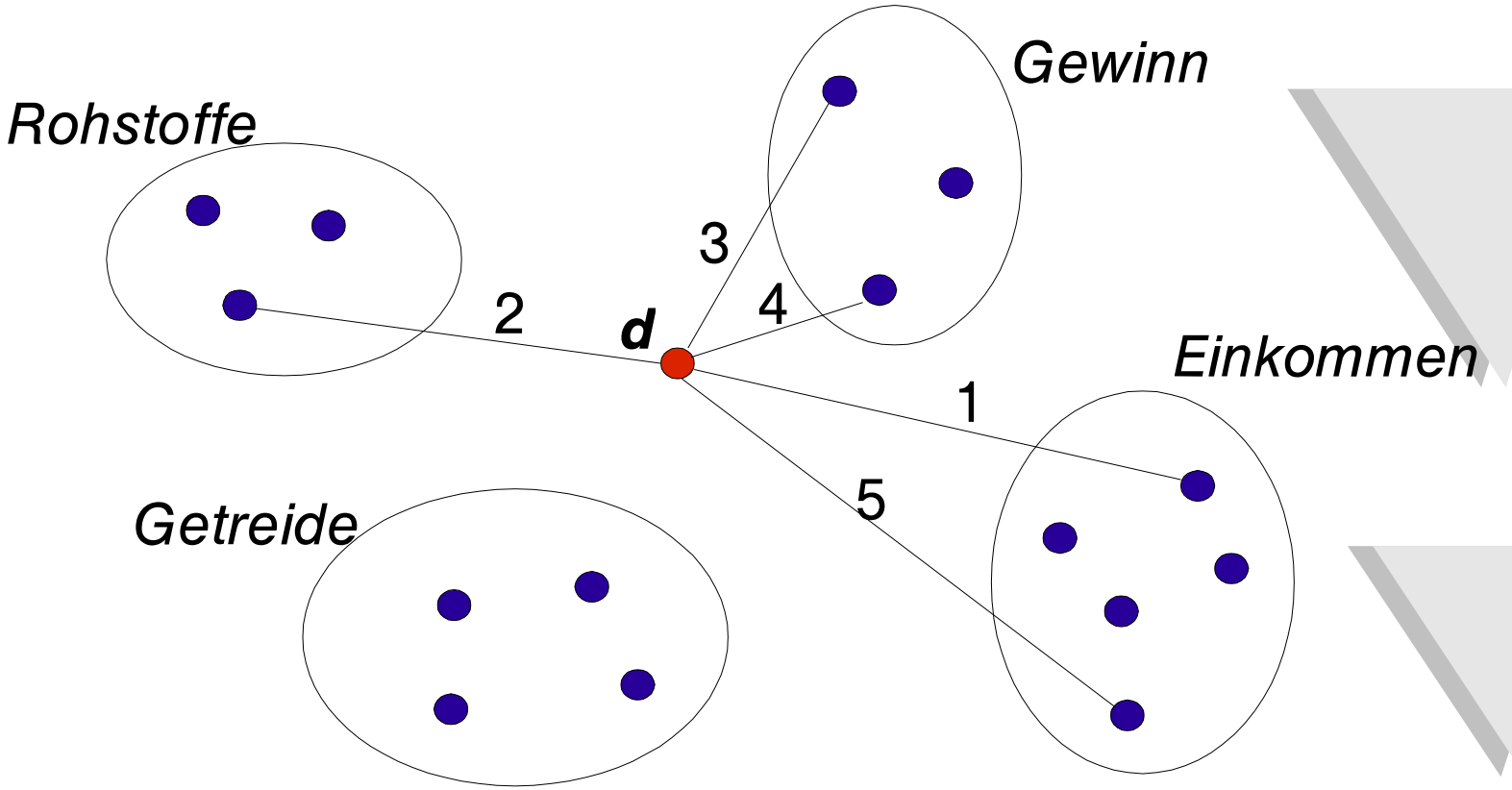
Text Mining

- Der kNN-Algorithmus
  - Trainingsphase



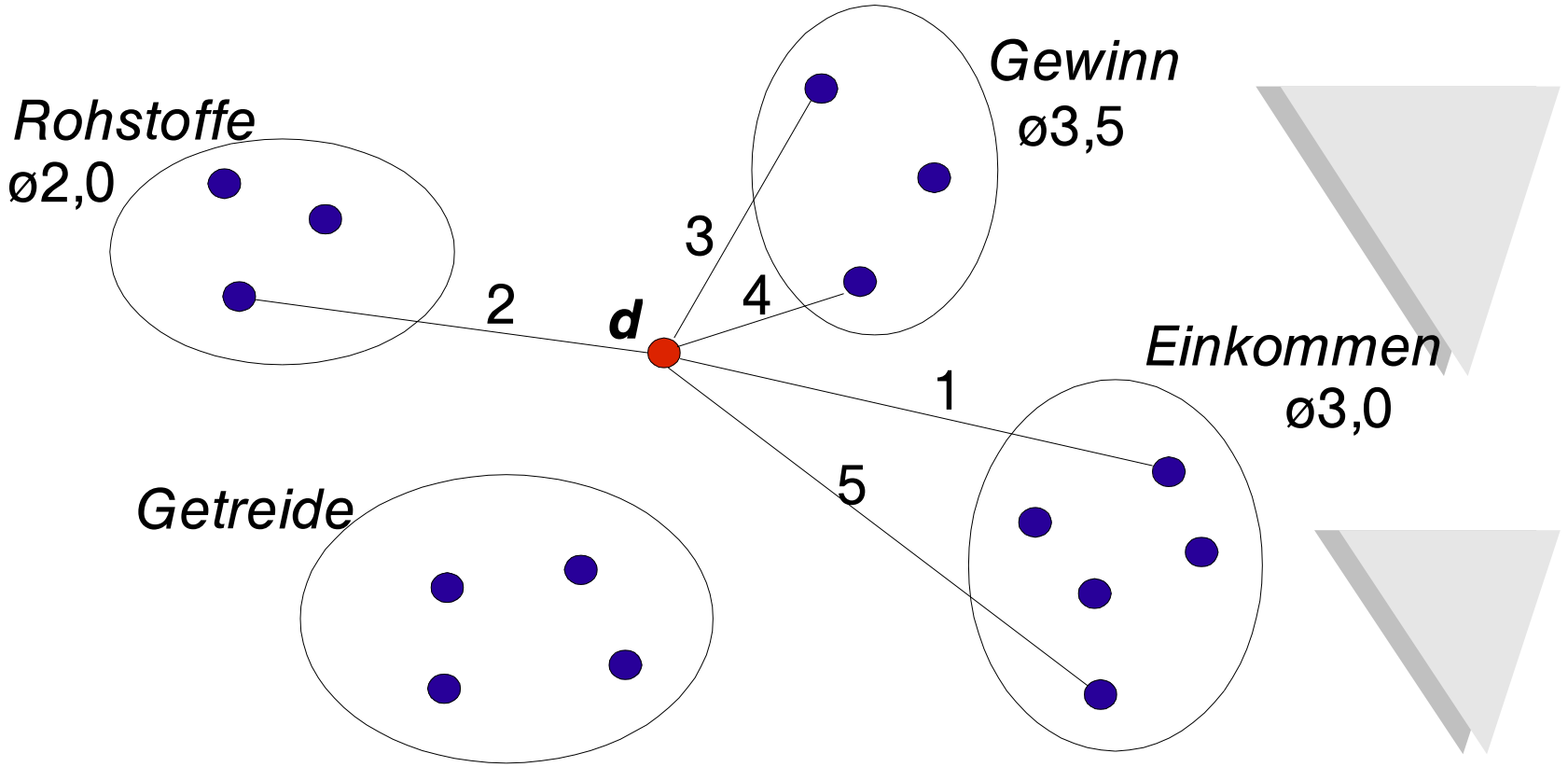
Text Mining

- Der kNN-Algorithmus
  - Finden der k-ähnlichsten Nachbarn von ***d***



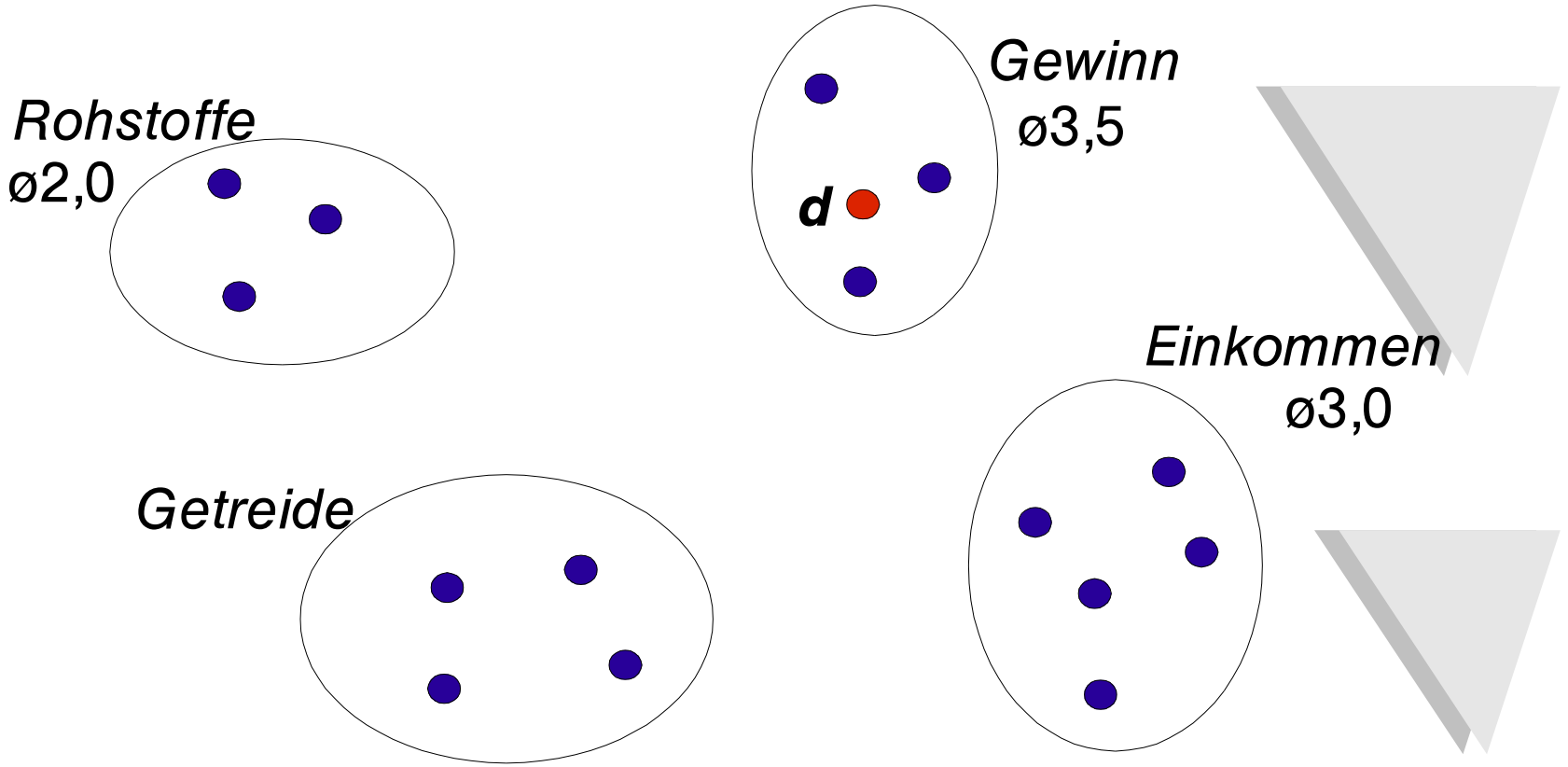
Text Mining

- Der kNN-Algorithmus
  - Gewichtung der Klassen



Text Mining

- Der kNN-Algorithmus
  - Klassenzuordnung

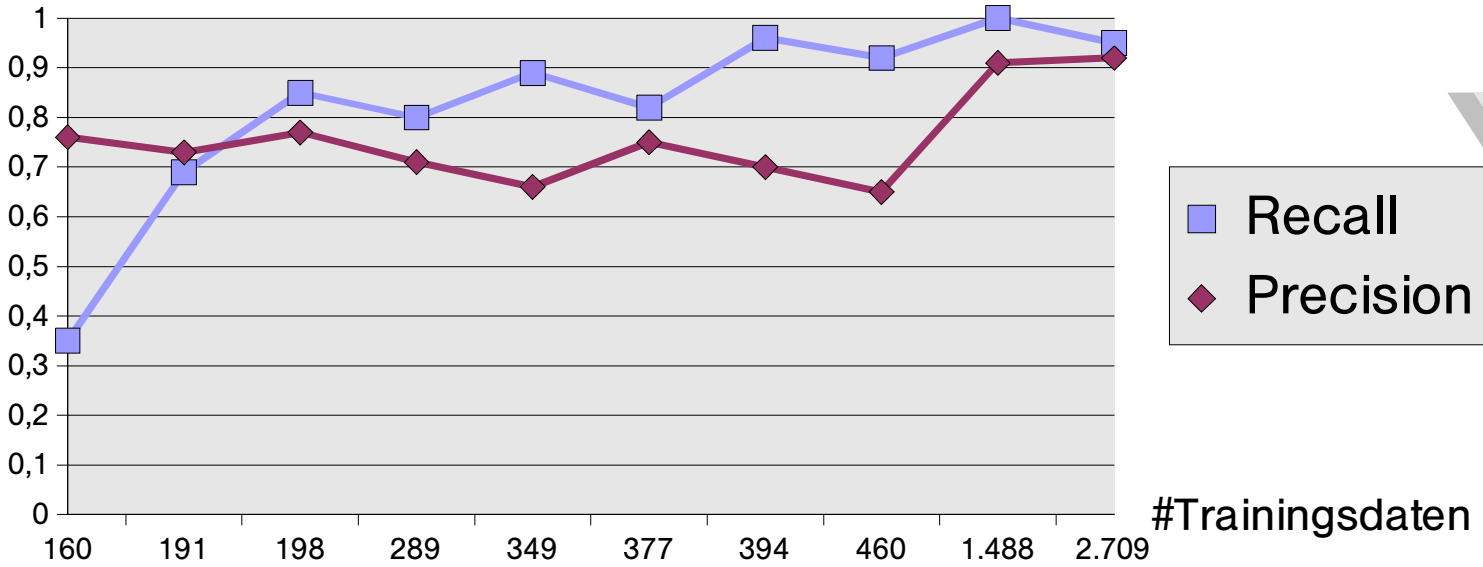


Text Mining

- Maßgrößen zur Effektivitätsmessung
  - Precision (Präzision) : $\frac{a}{(a+b)}$
  - Recall (Ausbeute) : $\frac{a}{(a+c)}$
    - a : Anzahl korrekt zugewiesener Dokumente
    - b: Anzahl der Dokumente, die der Kategorie fälschlicherweise zugewiesen wurden
    - c: Anzahl der Dokumente, die der Kategorie fälschlicherweise nicht zugewiesen wurden
  - Precision und Recall hängen voneinander ab

Text Mining

- Ergebnisse



- Je mehr Trainingsdaten, desto genauer das Ergebnis

Image Mining

- Ziel
  - Automatische Extraktion von semantisch aussagekräftigen Informationen
- Anwendungen
  - Medizin
    - Analyse von medizinischen Aufnahmen
  - Image-Retrieval
    - Kategorisieren von Bildern in „relevant“ und „irrelevant“ bzgl. einer Anwendung

Image Mining

- Ein informationsorientierter Image-Mining-Ansatz
  - Pixel-Ebene
    - Extrahieren potentiell relevanter Regionen
      - homogene Farbverteilung
      - Konturen
  - Objekt-Ebene
    - Identifizierung domänenspezifischer Merkmale der Regionen
      - Fläche, Länge, Form, ...
  - Semantische Ebene
    - Erkennen von Mustern in der identifizierten Objektmenge durch Anwenden eines Data-Mining-Algorithmus
  - Wissens-Ebene

Image Mining

- Anwendung bei der Erkennung von Tumorzellen auf medizinischen Aufnahmen
  - Aufgabe
    - Zählen von Tumorzellen zur Messung der Aktivität des Tumors
    - Hohe Treffergenauigkeit erforderlich
    - Reine Bildverarbeitungs-Algorithmen ungeeignet
    - Zählen von Tumorzellen durch Menschen zu zeitaufwendig und teuer
  - Ziel
    - Signifikante Steigerung der Genauigkeit durch Kombination von Bildverarbeitungs-Algorithmen und Image-Mining-Algorithmen

Image Mining

- Pixelebene
  - Erkennen potentiell relevanter Regionen
  - Problem bei medizinischen Aufnahmen
    - Kein homogener Hintergrund
      - Reflektionen
      - Überlappende Zellen, dadurch verschiedene Pixelintensitäten
    - Unklare Konturen

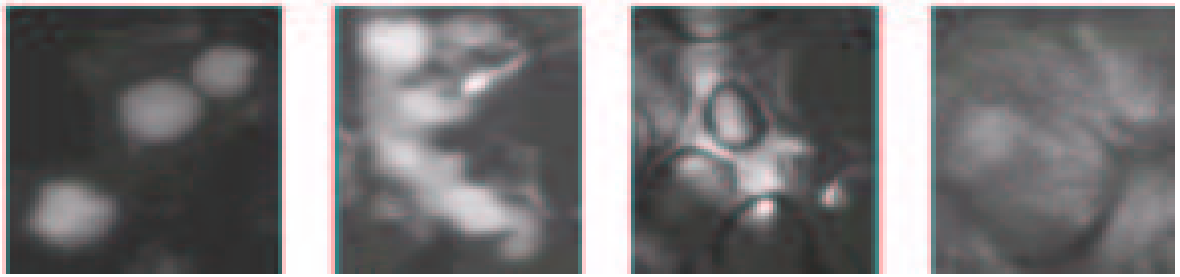


Image Mining

- Ansatz zur Extraktion von potentiell relevanten Objekten
  - Festlegen einer Pixelintensität, ab der eine Region als potentiell relevant deklariert wird
- Problem
  - Tumorzellen können in einem Teil des Bildes dunkler sein, als der Hintergrund in einem anderen Bildbereich
    - Es könnten zu viele oder zu wenige Tumorzellen erkannt werden, je nach Wahl des Schwellwertes

Image Mining

- Festlegen lokaler Schwellwerte
  - Ergebnis nach Entfernen des Hintergrundes

- Der „Water-Immersion-Algorithmus“

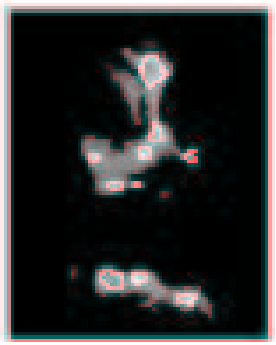


Image Mining

- Objektebene
  - Berechnung folgender Merkmale für jede erkannte Region
    - Fläche: Anzahl Pixel
    - Rundheit $= \frac{4 \cdot \pi \cdot \text{Fläche}}{\text{Umfang}^2}$
    - Dehnung $= \frac{L_{\text{Hauptachse}}}{L_{\text{Nebenachse}}}$
  - Jede Region wird durch diese drei Merkmale charakterisiert

Image Mining

- Semantische Ebene
  - Anlegen einer Trainingsmenge mit Ausprägungen der drei Merkmale, die auf einen Tumor schließen lassen
  - Anwenden einen Klassifizierungsalgorithmus
    - CBA
    - C4.5
    - Bayesche Klassifikation
  - Ergebnis
    - Ohne Data Mining: Fehlerrate von 40,1%
    - Mit Data Mining: Fehlerrate von 20%
    - Kombination aller drei Klassifizierungsalgorithmen: Fehlerrate von 18,7%

Video Mining

- Ziel
  - Organisieren von Videodaten derart, dass bisher unbekannte Informationen extrahiert werden können
- Video-Mining-Anwendungen
  - Verkehr
  - Medizin
  - Biologie

Video Mining

- Video-Mining in der Verkehrsüberwachung
  - Ziele
    - Erkennen von Staus
    - Erkennen von Unfällen
    - Analyse der Verkehrsbelastung
  - Eigenschaften von Überwachungsvideos
    - Meist stationäre Kameras
      - Gleichbleibender Hingergrund
    - Keine oder wenige Schnitte
    - Bekannte Objekte

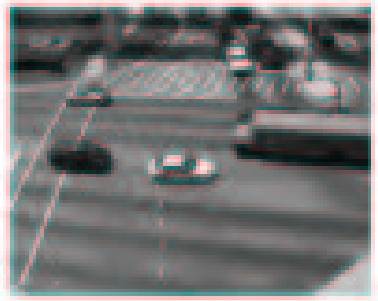
Video Mining

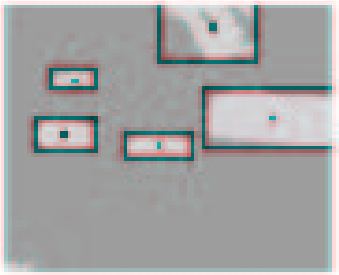
- Analyse des Verkehrsaufkommens an einer Kreuzung
  - (1) Entfernen des Hintergrundes
    - Wie beim Image-Mining
      - Zu aufwendig, da für jeden Frame durchzuführen
    - Extraktion von sich nicht bewegenden Objekten
    - Erzeugen eines Referenzframes in Form einer Referenzaufnahme der leeren Kreuzung
      - Schnelles Verfahren
      - Probleme bei wechselnden Lichtverhältnissen o.ä.

Video Mining

(2) Erkennen von Fahrzeugen

- Repräsentation der Fahrzeuge durch Rechtecke



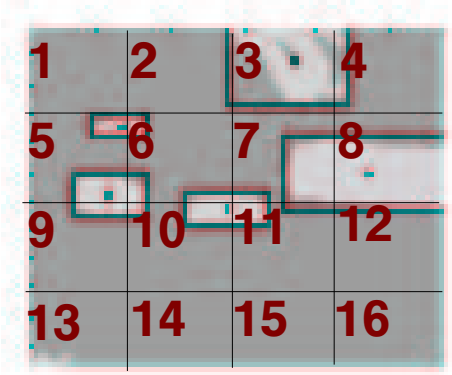


- Probleme
  - Verdeckte Fahrzeuge
  - Von oben kommende Fahrzeuge werden als ein langes Fahrzeug erkannt
- Zunächst als ein Fahrzeug behandeln. Eventuell später den möglichen Fahrzeugweg zurückverfolgen

Video Mining

(3) Verfolgen von Fahrzeugen

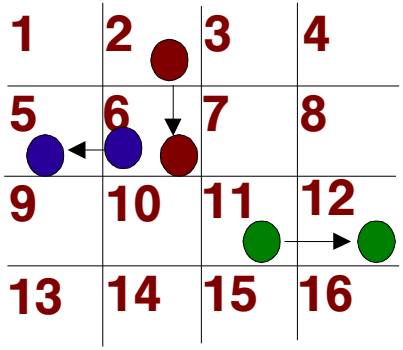
- Verbinden von zusammengehörigen Objekten zweier aufeinanderfolgender Frames
  - Jeweils zwei Objekte mit geringstem Abstand
  - Größe der Objekte mit einbeziehen
- Aufteilung jedes Frames in beschriftete Segmente



Video Mining

(4) Ergebnisse

- Multimedia Input Strings



A6 & B2 & C11
A5 & B2 & C12
A5 & B6
.....

- Ermöglicht Beantwortung von Fragen wie
  - Wie viele Fahrzeuge sind von links in die Kreuzung eingefahren und haben sie nach oben wieder verlassen?
  - Wie hoch war das Verkehrsaufkommen zwischen 08:00 und 08:30 ?

Mining in räumlichen Daten

- Eine räumliche Datenbank
  - Räumliche Objekte (Straßen, Häuser, Flüsse,...)
    - Stehen in Relation zu anderen Objekten in der Datenbank
      - Topologie (A berührt B, A liegt in B,...)
      - Distanz
      - Richtung
  - Nicht-räumliche Attribute je Objekt (Hausnummer, Wohnfläche, Bewohnerzahl, ...)
- Repräsentation mittels Nachbarschaftsgraphen
  - Knotenmenge
    - Menge der räumlichen Objekte
  - Kantenmenge
    - Knotenpaare, für die eine der Nachbarschafts-Relationen gilt

Mining in räumlichen Daten

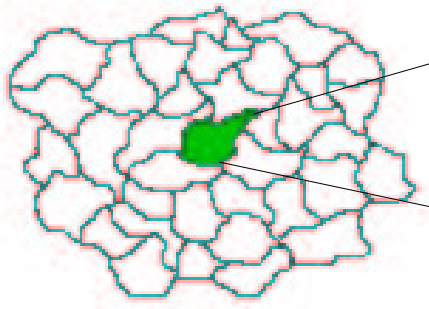
- Ziel
  - Erkennen von Mustern und Beziehungen zwischen
    - Räumlichen Objekten untereinander
      - Parks liegen häufig in der Nähe von Schulen
    - Räumlichen und nicht-räumlichen Daten
      - In der Nähe von Universitäten ist die Anzahl von Wohnungen pro Gebäude relativ hoch
- Traditionelle statistische Data-Mining-Methoden sind ungeeignet
  - Datensätze sind nicht unabhängig verteilt
  - Objekte beeinflussen sich gegenseitig

Mining in räumlichen Daten

- Finden von Assoziationen in räumlichen Daten
  - Korrelationen zwischen verschiedenen Charakteristiken in bestimmten Regionen
    - Beispiel
      - Regionen mit einer hohen Anzahl von Rentnern liegen meist in der Nähe von Gebirgen und Flüssen
- Clustering
  - Dichtebasiertes Clustering
    - Jedes Objekt innerhalb eines Clusters muss eine minimale Anzahl von Knoten in seiner Nachbarschaft haben

Mining in räumlichen Daten

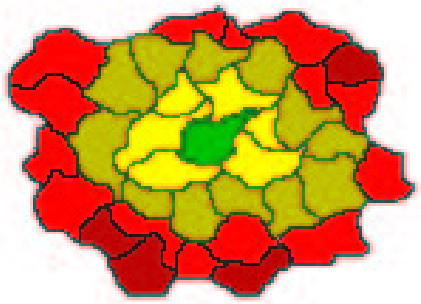
- Erkennen von Trends und Trendabweichungen
  - Analysieren von Veränderungen nicht-räumlicher Attribute bei der Entfernung von einem räumlichen Objekt
    - Festlegen eines Zentrums
      - Zeichnet sich durch besondere Merkmalsausprägung mind. eines nicht-räumlicher Attributes aus



Fläche	10 km
Einwohner	500
Arbeitslosenquote	1,00%
Rentneranteil	7,00%
....	

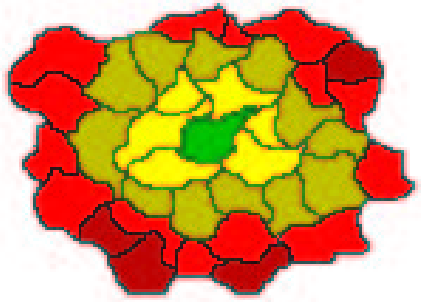
Mining in räumlichen Daten

- Berechnung eines theoretischen Trends
  - Annahme: Die Arbeitslosenquote steigt mit wachsender Enfternung vom Zentrum

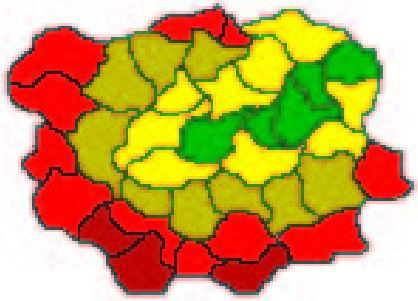


Mining in räumlichen Daten

- Berechnung eines theoretischen Trends
  - Annahme: Die Arbeitslosenquote steigt mit wachsender Enfternung vom Zentrum

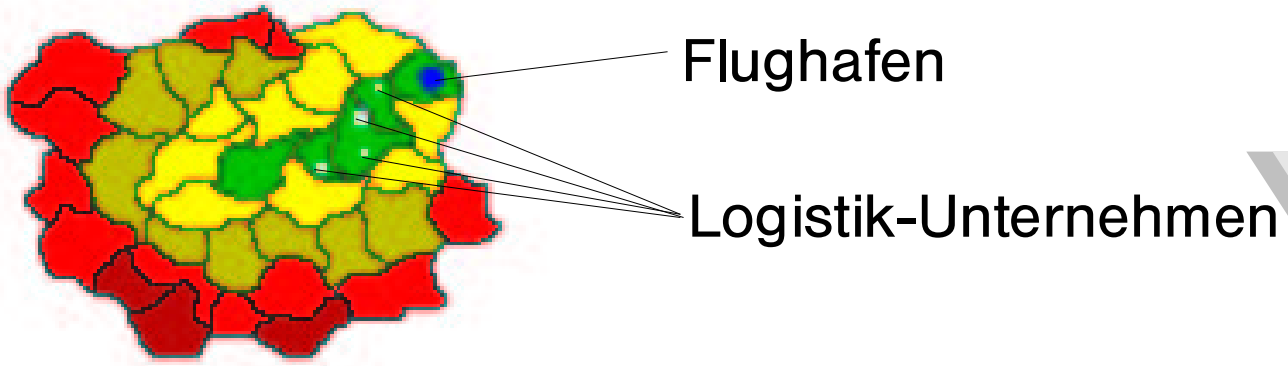


- Suche nach Trendabweichungen



Mining in räumlichen Daten

- Begründen der Abweichung



- Weitere Anwendungen
  - Verbrechensanalyse
  - Suche nach Ursachen für Naturkatastrophen

Soziale Auswirkungen

- Data Mining
  - Generieren von Informationen, die nicht offensichtlich sind
    - Vorteile
      - Vorhersagen von Naturkatastrophen
      - Unterstützung bei der Diagnose von Krankheiten
      - Personalisierte Produkte, personalisiertes Marketing
    - Nachteil
      - Bei der Bekanntgabe von Informationen ist von vornherein nicht bekannt, wieviel Wissen man tatsächlich preisgegeben hat

Trends im Data Mining

- Exponentielles Wachstum von Informationen
  - Vervielfältigung der Anwendungsmöglichkeiten
  - Neue Methoden zum Mining von komplexen Datentypen
  - Skalierbare Data-Mining-Methoden
  - Spezialisierte Mining-Methoden

Zusammenfassung

- Text Mining
  - Klassifizieren von Text-Dokumenten
    - Preprocessing, Indexierung, Dimensionsreduktion, Clustering
- Image Mining
  - Erkennen von Tumoren
    - Pixelebene, Objektebene, Semantische Ebene, Wissensebene
- Video Mining
  - Der Video-Mining-Prozess
    - Verkehrsüberwachung: Objektverfolgung, Multimedia Input Strings

Zusammenfassung

- Mining in räumlichen Daten
  - Erkennen von Trends und Trendabweichungen
  - Assoziationen in räumlichen Daten
- Soziale Auswirkungen und Trends im Data Mining