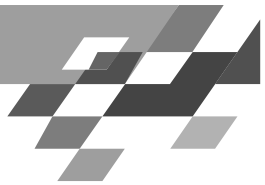
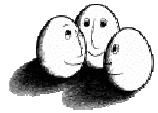


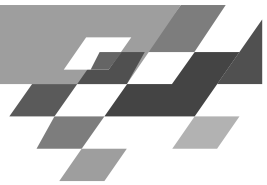
Texte als Daten

Web Mining

Textklassifikation

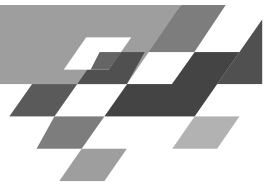
Verwendung des Modells für Textklassifikation für zeitgestempelte Daten





World Wide Web

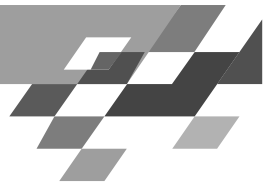
- Seit 1993 wächst die Anzahl der Dokumente -- geschätzt 12,9 Milliarden Seiten (2005)
- Ständig wechselnder Inhalt ohne Kontrolle, Pflege
  - Neue URLs
  - Neue Inhalte
  - URLs verschwinden
  - Inhalte werden verschoben oder gelöscht
- Verweisstruktur der Seiten untereinander
- Verschiedene Sprachen
- Unstrukturierte Daten



Informationsextraktion

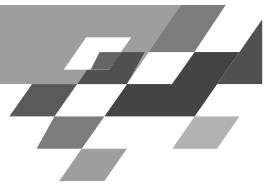
- Textstücke innerhalb der Dokumente finden
- Semantic Web: Auszeichnungssprache für Dokumente zur Verschlagwortung von Text(-teilen) durch Autoren
- Automatic tagging
- Named Entity Recognition (NER)

Machen wir jetzt nicht!



Aufgaben

- Indexierung möglichst vieler Seiten (Google)
- Suche nach Dokumenten, ranking der Ergebnisse z.B. nach Häufigkeit der Verweise auf das Dokument (PageLink -- Google)
- Kategorisierung (Klassifikation) der Seiten manuell (Yahoo), automatisch
- Strukturierung von Dokumentkollektionen (Clustering)
- Personalisierung:
  - Navigation durch das Web an Benutzer anpassen
  - Ranking der Suchergebnisse an Benutzer anpassen



Information Retrieval

- Ein Dokument besteht aus einer Menge von Termen (Wörtern).
  - Bag of words: Vektor, dessen Komponenten die Häufigkeit eines Wortes im Dokument angeben.
- Für alle Dokumente gibt es eine Termliste mit Verweis auf die Dokumente.
  - Anzahl der Dokumente, in denen das Wort vorkommt.

Beispiel zur Klassifikation

To: rueping@ls8.cs.uni-dortmund.de

Subject: **Astonishing** Guaranteed ~~**XXX**~~ **Pictures FREE**! Gao

In the next 2 minutes you are going **to** learn how **to** get access **to** totally **FREE xxx pictures**. Let me show you the secrets I have learned **to** get **FREE porn** passwords. Indeed, with this **in** mind lets take a quick look below **to** see what you get, ok?

1	astonishing
3	free
2	in
	⋮
2	pictures
1	porn
0	SVM
5	to
0	university
2	XXX

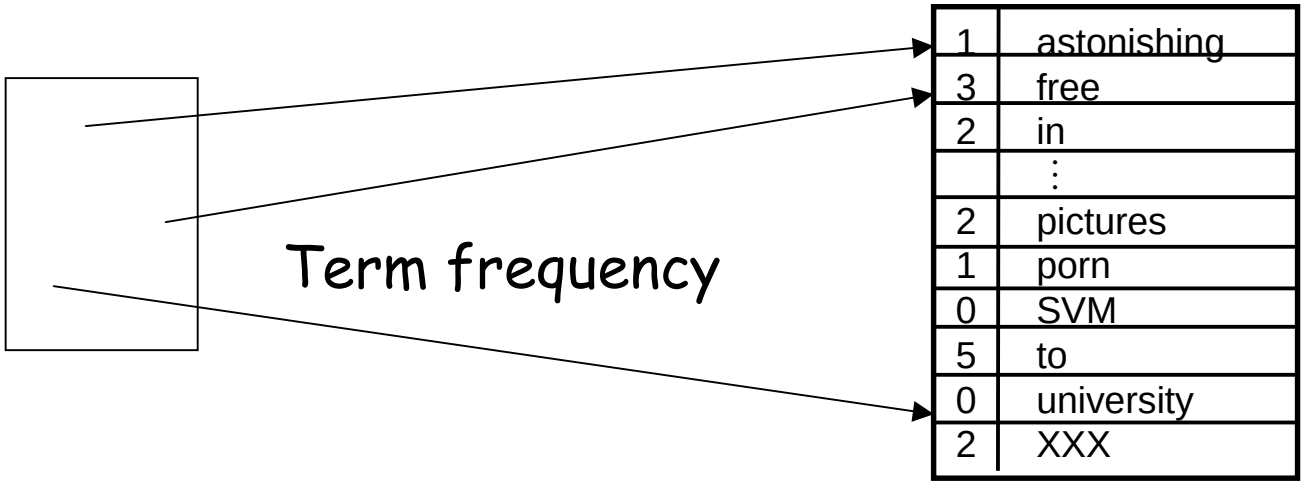
*

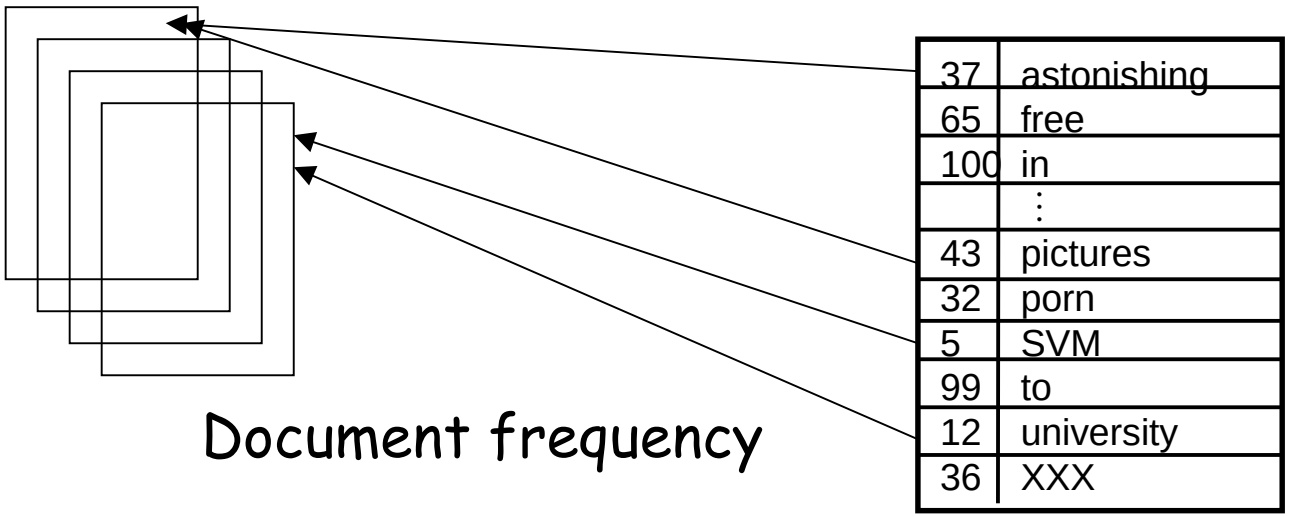
0.1
0.4
0.0
⋮
0.2
1.1
-0.6
0.0
-0.4
0.9

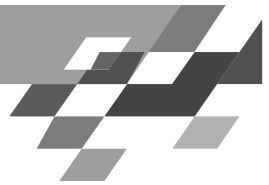
SVM

> 0

Texte als Daten







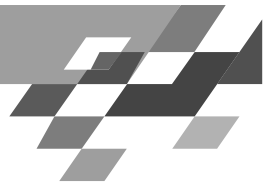
TFIDF

- Term Frequenz: wie häufig kommt ein Wort w_i in einem Dokument d vor? TF(w_i,d)
- Dokumentenfrequenz: in wie vielen Dokumenten einer Kollektion D kommt ein Wort w_i vor? DF(w_i)
- Inverse Dokumentenfrequenz:

$$IDF(D, w_i) = \log \frac{|D|}{DF(w_i)}$$

- Bewährte Repräsentation:

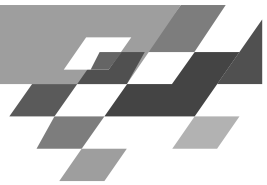
$$TFIDF(w_i, D) = \frac{TF(w_i, d)IDF(w_i, D)}{\sqrt{\sum_j \left[TF(w_j, d)IDF(w_j, D)\right]^2}}$$



Textklassifikation

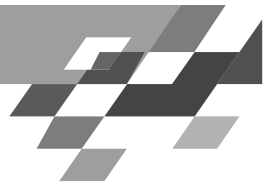
- Thorsten Joachims „The Maximum-Margin Approach to Learning Text Classifiers“ Kluwer 2001
- Modell der Textklassifikation TCat
- Verbindung zur SVM-Theorie

➔ theoretisch begründete Performanzabschätzung



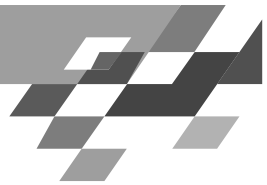
Eigenschaften der Textklassifikation1

- Hochdimensionaler Merkmalsraum
  - Reuters Datensatz mit 9 603 Dokumenten: V=27 658 verschiedene Wörter
  - Heapes Gesetz: Anzahl aller Wörter (s)
    $V = k\, s^{\beta}$
  - Beispiel:
    - Konkatenieren von 10 000 Dokumenten mit je 50 Wörtern zu einem,
    - K=15 und β=0,5
    - ergibt V=35 000 – stimmt!



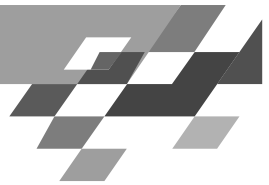
Eigenschaften der Textklassifikation2

- Heterogener Wortgebrauch
  - Dokumente der selben Klasse haben manchmal nur Stoppwörter gemeinsam!
  - Es gibt keine relevanten Terme, die in allen positiven Beispielen vorkommen.
  - Familienähnlichkeit (Wittgenstein): A und B haben ähnliche Nasen, B und C haben ähnliche Ohren und Stirn, A und C haben ähnliche Augen.



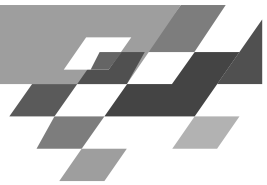
Eigenschaften der Textklassifikation3

- Redundanz der Merkmale
  - Ein Dokument enthält mehrere die Klasse anzeigende Wörter.
  - Experiment:
    - Ranking der Wörter nach ihrer Korrelation mit der Klasse.
    - Trainieren von Naive Bayes für Merkmale von Rang
      1 - 200 (90% precision/recall),
      201 - 500 (75%)
      601 - 1000 (63%)
      1001- 2000 (59%)
      2001- 4000 (57%)
      4001- 9947 (51%) -- zufällige Klassifikation (22%)



Eigenschaften der Textklassifikation4

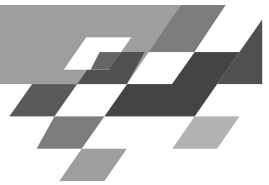
- Dünn besetzte Vektoren
  - Reuters Dokumente durchschnittlich 152 Wörter lang
    - mit 74 verschiedenen Wörtern
    - bei den meisten Wörtern 0
  - Euklidsche Länge der Vektoren klein!



Eigenschaften der Textklassifikation5

- Zipfs Gesetz: Verteilung von Wörtern in Dokumentkollektionen ist ziemlich stabil.
  - Ranking der Wörter nach Häufigkeit (r)
  - Häufigkeit des häufigsten Wortes (max)
  - 1/r max häufig kommt ein Wort des Rangs r vor.
- Generalisierte Verteilung von Häufigkeit nach Rang (Mandelbrot): c ist Größe der Dokumentkollektion in Wortvorkommen

$$\frac{c}{(k+r)^{\phi}}$$

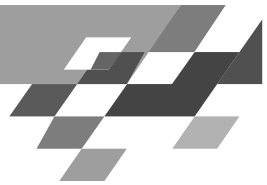


Plausibilität guter Textklassifikation durch SVM

- R sei Radius des Balles, der die Daten enthält. Dokumente werden auf einheitliche Länge normiert, so dass R=1.
- Margin sei δ, so dass großes δ kleinem R^2/δ^2 entspricht

Reuters	R^2/δ^2	$\sum_{i=1}^{n} \xi_i$
Earn	1143	0
acquisition	1848	0
money-fx	1489	27
grain	585	0
crude	810	4

Reuters	R^2/δ^2	$\sum_{i=1}^{n} \xi_i$
trade	869	9
interest	2082	33
ship	458	0
wheat	405	2
corn	378	0



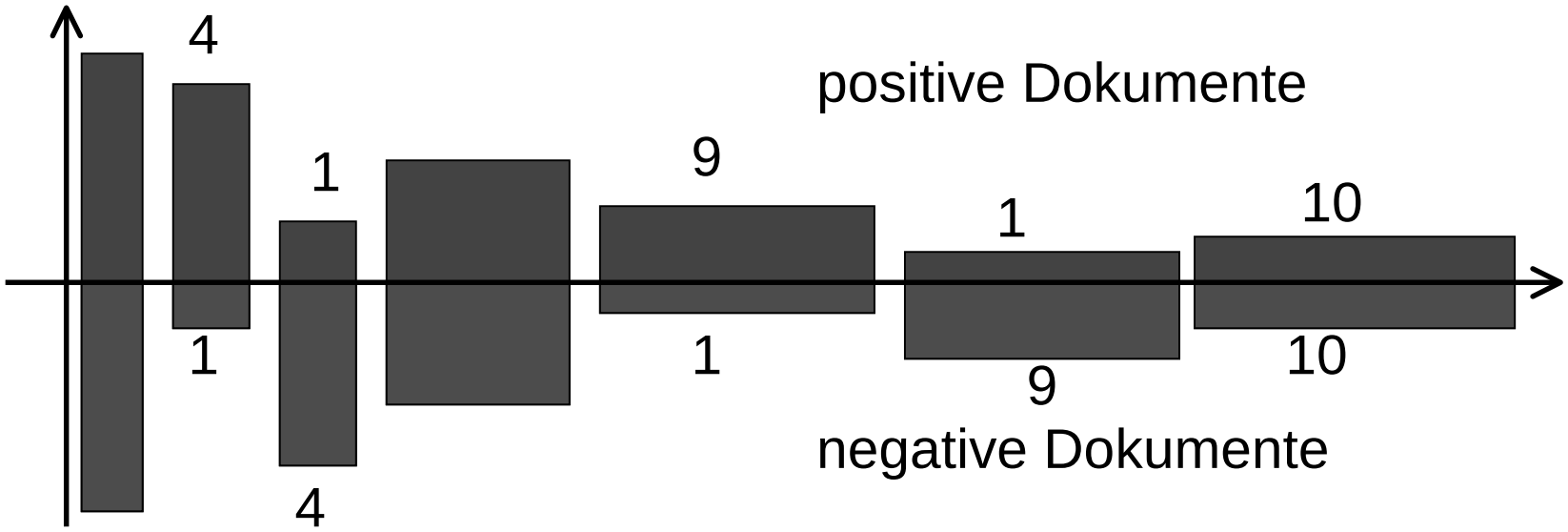
TCat Modell -- Prototyp

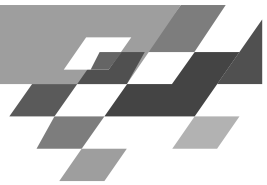
- Hochdimensionaler Raum: 11100 Wörter im Lexikon
- Dünn besetzt: Jedes Dokument hat nur 50 Wörter, also mindestens 11050 Nullen
- Redundanz: Es gibt 4 mittelhäufige und 9 seltene Wörter, die die Klasse anzeigen
- Verteilung der Worthäufigkeit nach Zipf/Mandelbrot.
- Linear separierbar mit b=0,
  w= 0,23 für mittelhäufige Wörter in POS,
  w= -0,23 für mittelhäufige Wörter in NEG,
  w= 0,04 für seltene Wörter in POS,
  w= -0,04 für seltene Wörter in NEG,
  w=0 sonst

$$\sum_{i=1}^{11100} w_i x_i$$

TCat im Bild

- 20 aus 100 Stoppwörtern, 5 aus 600 mittelhäufigen und 10 aus seltenen Wörtern kommen in POS- und NEG-Dokumenten vor;
  4 aus 200 mittelhäufigen Wörtern in POS, 1 in NEG,
  9 aus 3000 seltenen Wörtern in POS, 1 in NEG
  (Es müssen nicht immer die selben Wörter sein!)



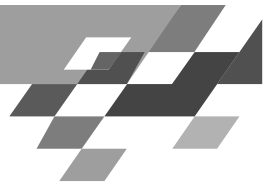


TCat

The TCat concept

$$TCat([p_1 : n_1 : f_1], \ldots, [p_s : n_s : f_s])$$

describes a binary classification task with s sets of disjoint features. The i-th set includes f_i features. Each positive example contains p_i occurences of features from the respective set and each negative example contains n_i occurrences. The same feature can occur multiple times in one document.
(Joachims 2002)

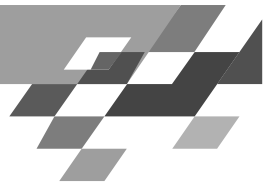


TCat zum Bild

TCat([20 :20: 100] sehr häufig

[4: 1: 200] [1: 4: 200] [5: 5: 600]
mittel häufig

[9: 1: 3000] [1: 9: 3000] [10 : 10: 4000]
selten
)



Lernbarkeit von TCat durch SVM

(Joachims 2002) Der erwartete Fehler einer SVM ist nach oben beschränkt durch:

$$\frac{R^2}{n+1}\frac{a+2b+c}{ac-b^2}$$

$$a=\sum_{i=1}^{s}\frac{p_i^2}{f_i}$$

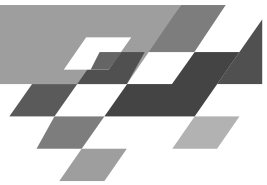
$$b=\sum_{i=1}^{s}\frac{p_i^2 n_i}{f_i}$$

$$c=\sum_{i=1}^{s}\frac{n_i^2}{f_i}$$

$$R^2=\sum_{r=1}^{d}\left(\frac{c}{(r+k)^{\phi}}\right)^2$$

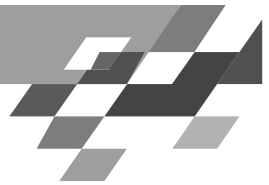
Es gibt l Wörter, s Merkmalsmengen, für einige i: $p_i \neq n_i$ und die Termhäufigkeit befolgt Zipfs Gesetz. Wähle d so, dass:

$$\sum_{r=1}^{d}\frac{c}{(r+k)^{\phi}}=l$$

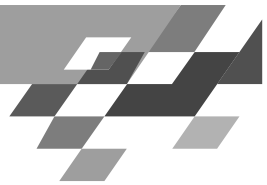


Was wissen Sie jetzt?

- Die automatische Klassifikation von Texten ist durch das WWW besonders wichtig geworden.
- Texte können als Wortvektoren mit TFIDF dargestellt werden. Die Formel für TFIDF können Sie auch!
- Textkollektionen haben bzgl. der Klassifikation die Eigenschaften: hochdimensional, dünn besetzt, heterogen, redundant, Zipfs Gesetz.
- Sie sind mit breitem margin linear trennbar.
- Das TCat-Modell kann zur Beschränkung des erwarteten Fehlers eingesetzt werden. Die Definition von TCat kennen Sie mindestens, besser wäre noch die Fehlerschranke zu kennen.

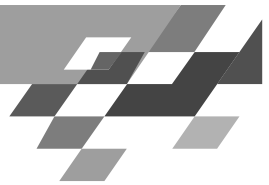


Und jetzt wenden wir das Gelernte auf ein Gebiet fernab von Texten an!



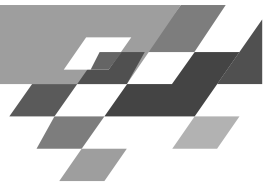
Lokale Muster

- Lokale Muster beschreiben seltene Ereignisse.
- Gegeben ein Datensatz, für den ein globales Modell bestimmt wurde, weichen lokale Muster davon ab.
  - Lokale Muster beschreiben Daten mit einer internen Struktur, z.B. Redundanz, Heterogenität



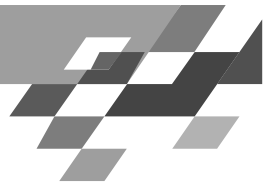
Zeit-gestempelte Daten

- Zeit-gestempelte Daten können transformiert werden in:
  - Eine Menge von of Ereignissen,
  - Zeitintervalle,
  - Zeitreihen.
- Aufgaben sind
  - Vorhersage von Ereignissen (Winepi),
  - Entdeckung von Relationen zwischen Intervallen (Höppner),
  - Klassifikation von Prozessen.



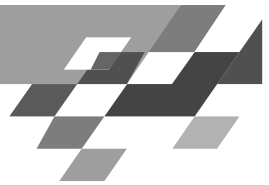
Klassische Methoden

- Zeitreihenanalyse für Vorhersage, Trend und Zyklus Erkennung
- Indexing und clustering von Zeitreihen (time warping)
- Segmentierung (motif detection)
- Entdeckung von Episoden
  - frequent sets,
  - chain logic programs (grammars)
- Regression



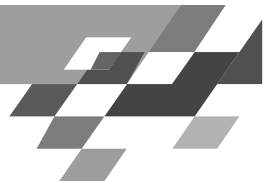
Beispielrepräsentation

- Die Beispielrepräsentation L_E bestimmt die Anwendbarkeit der Methoden.
- Bedeutung von L_E lange unterschätzt.
- Suche nach gutem L_E ist aufwändig.
- Transformieren der Rohdaten in L_E auch.



Einige Repräsentationen L_E für zeitgestempelte Daten

- Schnappschuss: ignoriere Zeit, nimm nur den aktuellen Zustand.
- Ereignisse mit Zeitintervallen: aggregiere Zeitpunkte zu Intervallen, wende frequent set mining an.
- Generierte Merkmale: hier: transformiere Zeitinformation in Häufigkeitsmerkmale!



Häufigkeitsmerkmale für Zeitaspekte

- Term frequency: wie oft änderte Attribut *A* seinen Wert a_i für ein Objekt c_j.

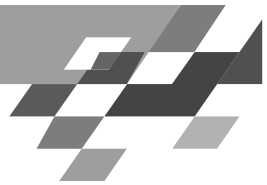
$$tf(a_i, c_j) = \left\| \left\{ x \in timepoints \middle| a_i of\ c_j\ changed \right\} \right\|$$

- Document frequency: in wie vielen Objekten c_j änderte Attribut *A* seinen Wert a_i.

$$df(a_i) = \left\| \left\{ c_j \in C \middle| a_i\ of\ c_j\ changed \right\} \right\|$$

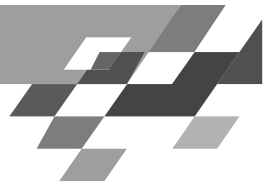
- TF/IDF:

$$tfidf(a_i) = tf(a_i, c_j) \log \frac{\|C\|}{df(a_i)}$$



Fallstudie SwissLife

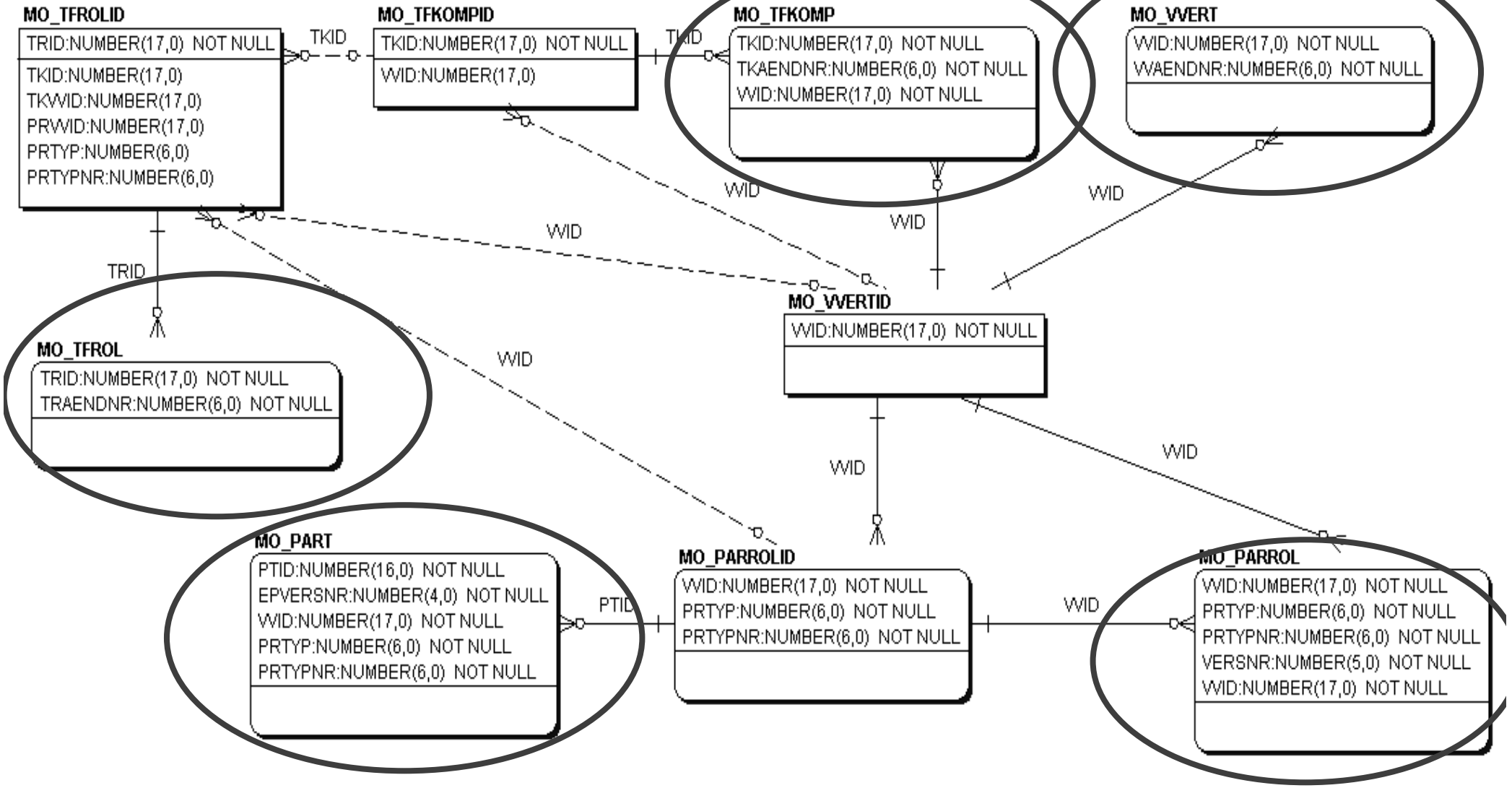
- Lokale Muster
  - Seltenes Ereignis der Kündigung
  - Lokales Muster weicht ab vom generellen Modell
  - Interne Struktur in lokalen Mustern
- Zeit-gestempelte Daten
  - Schnappschuss
  - Zeitintervall
  - Generierte Merkmale: TFIDF

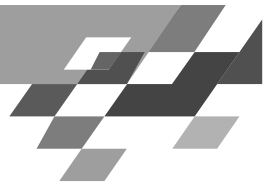


Lokale Muster in Versicherungsdaten

- Nur 7.7% der Verträge enden vorzeitig (customer churn).
- Für einige Attribute weicht die likelihood in der churn-Klasse von der globalen ab.
- Interne Struktur:
  - Überlappung: häufige Mengen in churn Verträgen sind auch häufig in fortgesetzten Verträgen.
  - Redundanz: in jedem Vertrag gibt es mehrere Attribute, die auf Fortsetzung oder Kündigung hinweisen.
  - Heterogenität: Es gibt gekündigte Verträge, die nicht ein einziges Attribut gemeinsam haben.

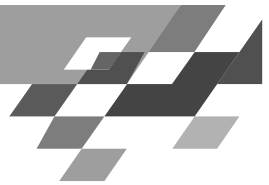
Database


MO_TFROLID
TRID:NUMBER(17,0) NOT NULL
TKID:NUMBER(17,0)
TKWID:NUMBER(17,0)
PRWID:NUMBER(17,0)
PRTYP:NUMBER(6,0)
PRTYPNR:NUMBER(6,0)
TKID
MO_TFKOMPID
TKID:NUMBER(17,0) NOT NULL
WID:NUMBER(17,0)
TKID
MO_TFKOMP
TKID:NUMBER(17,0) NOT NULL
TKAENDNR:NUMBER(6,0) NOT NULL
WID:NUMBER(17,0) NOT NULL
MO_WVERT
WID:NUMBER(17,0) NOT NULL
WAENDNR:NUMBER(6,0) NOT NULL
WID
WID
WID
WID
TRID
MO_TFROL
TRID:NUMBER(17,0) NOT NULL
TRAENDNR:NUMBER(6,0) NOT NULL
MO_WVERTID
WID:NUMBER(17,0) NOT NULL
WID
WID
WID
MO_PART
PTID:NUMBER(16,0) NOT NULL
EPVERSNR:NUMBER(4,0) NOT NULL
WID:NUMBER(17,0) NOT NULL
PRTYP:NUMBER(6,0) NOT NULL
PRTYPNR:NUMBER(6,0) NOT NULL
PTID
MO_PARROLID
WID:NUMBER(17,0) NOT NULL
PRTYP:NUMBER(6,0) NOT NULL
PRTYPNR:NUMBER(6,0) NOT NULL
WID
MO_PARROL
WID:NUMBER(17,0) NOT NULL
PRTYP:NUMBER(6,0) NOT NULL
PRTYPNR:NUMBER(6,0) NOT NULL
VERSNR:NUMBER(5,0) NOT NULL
WID:NUMBER(17,0) NOT NULL



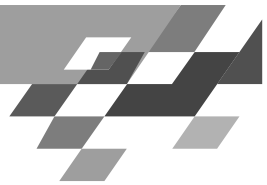
Contract Table

VVID	VVAENDNR	VVWIVON	VVWIBIS	VVAENDAT	VVAENDART	...
16423	1	1946	1998	1946	1000	
16423	2	1998	1998	1998	27	
16423	3	1998	1998	1998	4	
16423	4	1998	1998	1998	54	
16423	5	1998	1998	1998	4	
16423	6	1998	9999	1998	61	
5016	1	1997	1999	1997	33	
5016	2	1999	2001	1999	33	
5016	3	2001	2001	2001	33	
5016	4	2001	2001	2001	33	
5016	5	2001	2002	2001	81	
5016	6	2002	9999	2001	94	
...	...	...	...	...	...	...



Datensatz

- Tabellen enthalten Informationen über
  217586 Komponenten and
  163745 Kunden
- Attribute:
  - 14 Attributes ausgewählt
  - Eines der Attribute gibt den Grund an für einen Wechsel. Es gibt 121 Gründe. Daraus werden 121 Boolean Attribute.
  - 134 Attribute mit TFIDF Werten.



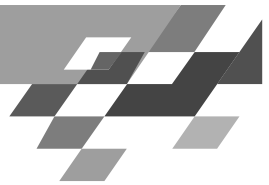
Erste Experimente

- Bei SwissLife wurde die Abweichung der Wahrscheinlichkeit bestimmter Attributwerte in gekündigten und fortgesetzten Verträgen festgestellt anhand der Schnappschuss-repräsentation $\Rightarrow$ keine operationale Vorhersage.
- Höppners Ansatz mit Apriori auf Zeitintervallen und ihren Relationen $\Rightarrow$ selbe Regeln gültig für Abbruch und Fortsetzung.

Calculating Term Frequency

VVID	...	VVSTACD	VVPRFIN	VVPRZA	VVINKZWEI	VVBEG	VVEND	VVINKPRL	...
16423		4	1	2	2	1946	1998	295,29	
16423		4	1	2	2	1946	1998	295,29	
16423		4	5	2	0	1946	2028	0	
16423		5	3	2	0	1946	2028	0	
16423		4	1	2	2	1946	1998	295,29	
16423		5	3	2	0	1946	1998	0	

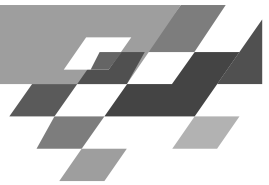
3	VVSTACD
4	VVPRFIN
0	VVPRZA
3	VVINKZWEI
0	VVBEG
2	VVEND
3	VVINKPRL



Experimente mit der TFIDF Repräsentation

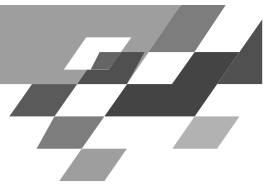
- Vergleich der originalen Repräsentation und der TFIDF
  - 10fold cross validation
    - Apriori mit Konklusion "churn"
    - J4.8
    - Naive Bayes
    - mySVM mit linearem Kern
  - F-measure balanciert precision und recall gleich.

Alle Lernalgorithmen werden besser mit der TFIDF- Repräsentation.



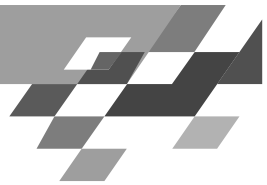
Resultate (F-measure)

Lerner	TF/IDF repr.	Original repr.
Apriori	63.35	30.24
J4.8	99.22	81.21
Naive Bayes	51.8	45.41
mySVM	97.95	16.06



Erklärung?

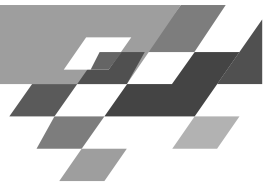
- TF/IDF stammt aus Lernen über Texten.
- Dazu gibt es eine Theorie -- TCat.
- Können wir die auch hier einsetzen??



Datenbeschreibung im TCat Modell

- Tcat ([2:0:2], [1:4:3], # high frequency
  [3:1:3], [0:1:4], # medium frequency
  [1:0:19], [0:1:64], # low frequency
  [1:1:39] # rest
)

[1:4:3]: 3 Merkmale kommen 1 mal in positiven und 4 mal in negativen Beispielen vor.



Learnability of TCat

- Error bound (Joachims 2002)

$$\frac{R^2}{n+1}\frac{a+2b+c}{ac-b^2}$$

$$a=\sum_{i=1}^{s}\frac{p_i^2}{f_i}$$ a= 5.41

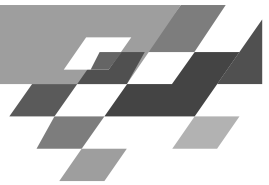
$$b=\sum_{i=1}^{s}\frac{p_i^2 n_i}{f_i}$$ b= 2.326

$$c=\sum_{i=1}^{s}\frac{n_i^2}{f_i}$$ c= 5.952

Nach 1000 Beispielen
erwarteter Fehler ≤ 2.2%

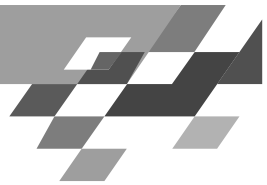
Tatsächlicher Fehler 2.05%

$$R^2=\sum_{r=1}^{d}\left(\frac{c}{(r+k)^{\phi}}\right)^2$$ $R^2 \leq 37$



Experimente zu lokalen Mustern

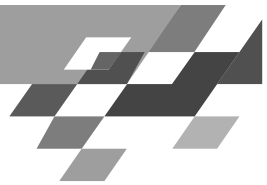
- Durch TCat-Konzepte Daten künstlich generieren.
- Lokale Muster als seltene Ereignisse mit interner Struktur.



Lokale Muster: Verzerrte Verteilung

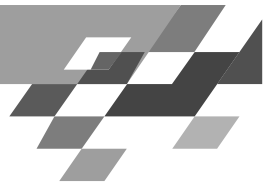
- 10 000 Beispiele mit 100 Attributen
- SVM runs mit 10 fold cross validation

L_E	Target concept:	Verzerrung:
TF/IDF	1. change of a particular attribute	50% 25% 12.5%
Boolean	2. frequency of changes	6.25%



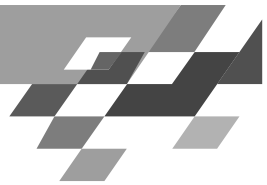
Lokale Muster: Strukturen

- 10 000 Beispiele mit 100 Attributen
- 20 Attribute wechseln pro Beispiel (dünn besetzt)
- Variieren:
  - Heterogenität: f_i/p_i Beispiele der selben Klasse haben kein gemeinsames Attribut {4, 5, 10, 20}
  - Redundanz: p_i/f_i oder n_i/f_i für die Redundanz innerhalb einer Klasse {0.5, 0.2, 0.1}
  - Überlappung: einige Attribute sind häufig in beiden Klassen {0.25, 0.66}



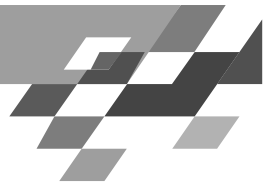
Resultate

- Für alle Kombinationen ohne Überlappung sind die Lernergbnisse 100% in Boolean und im TF/IDF- Format.
- Mehr Überlappung verschlechtert das Lernen bei Boolean auf 68.57% F-measure.
- Für alle Kombinationen (auch mit großer Überlappung) erreicht das Lernen mit TF/IDF Daten 100% precision und recall.



Navigation im L_E Raum

- Zunehmende Größe des Datensatzes:
  Schnappschuss < Intervalle < Boolean < TF/IDF
- TF/IDF ist günstig für lokale Muster, wenn diese Redundanz, Heterogenität als Struktur aufweisen.
- Berechnung des TCat Modells für gegebene Daten implementiert $\Rightarrow$ Fehlerschranke angebbar.
- Transformation der Rohdaten in TF/IDF implementiert.



Was wissen Sie jetzt?

- Lokale Muster haben manchmal die typische TCat-Struktur.
- Sie haben gesehen, wie manche zeitgestempelte Datenbanken in TCat-Modelle transformiert werden können.
- Die Lernbarkeit mit linearer SVM der so transformierten Daten können Sie ausrechnen.