Beispiel: Textklassifikation

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

Hochdimensionalität

- Hochdimensionaler Merkmalsraum:
  - 27 658 verschiedene Wörter in 9 603 Dokumenten des Reuters-Archivs
  - 38 359 verschiedene Wörter in 3 957 Dokumenten der WebKB-Sammlung
  - 38 679 verschiedene Wörter in 10 000 Dokumenten des Ohsumed-Archivs.
- Heap 1978: Anzahl verschiedener Wörter V ergibt sich bei s Wörtern in einem Dokument: $V=k\ s^{ß}$, wobei k in [10..100] und ß in [0,4..0,6]. Bei 10 000 Dokumenten à 50 Wörtern also $35000=15\ 500000^{0,5}$

Heterogenität

- Heterogenität: Dokumente derselben Klasse haben wenig Überschneidung bzgl. Der in ihnen vorkommenden Wörter. Es gibt sogar Dokumente, die derselben Kaegorie angehören und nicht ein einziges Inhaltswort gemeinsam haben!
- Familienähnlichkeit (Wittgenstein 1967)
- Selektion der wichtigsten Wörter (Merkmale) führt sofort zu schlechteren Lernergebnissen.

Redundanz

- Eine Menge von Wörtern kommt häufig gemeinsam vor innerhalb einer Textkategorie.
- Lässt man die besten Wörter weg (die mit der höchsten Korrelation zur Kategorie), ist immer noch ein statistischer Zusammenhang zwischen den Wörtern und der Textkategorie gegeben.
- Ordnet man alle Wörter nach der Korrelation mit der Kategorie in Ränge und nimmt nur die von Rang 4001 bis 9947, so ist ein naive Bayes Klassifikator weit besser als Zufall. (Joachims 2000 bzw. 2002)

Dünn besetzte Vektoren

- Jedes einzelne Dokument besitzt nur wenige Merkmale (verschiedene Wörter aus der Gesamtmenge). Die Euklidsche Länge ist kurz.
  - Reuter Dokumente sind 152 Wörter lang (im Durchschnitt) mit 74 verschiedenen Wörtern (von 27 658)
  - WebKB Dokumente sind 277 Wörter lang mit 130 verschiedenen Wörtern
  - Ohsumed Dokumente sind 209 Wörter lang, 100 Wörter verschieden.

TFIDF

- Term Frequenz: wie häufig kommt ein Wort w_i in einem Dokument d vor? TF(w_i,d)
- Dokumentenfrequenz: in wievielen 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:

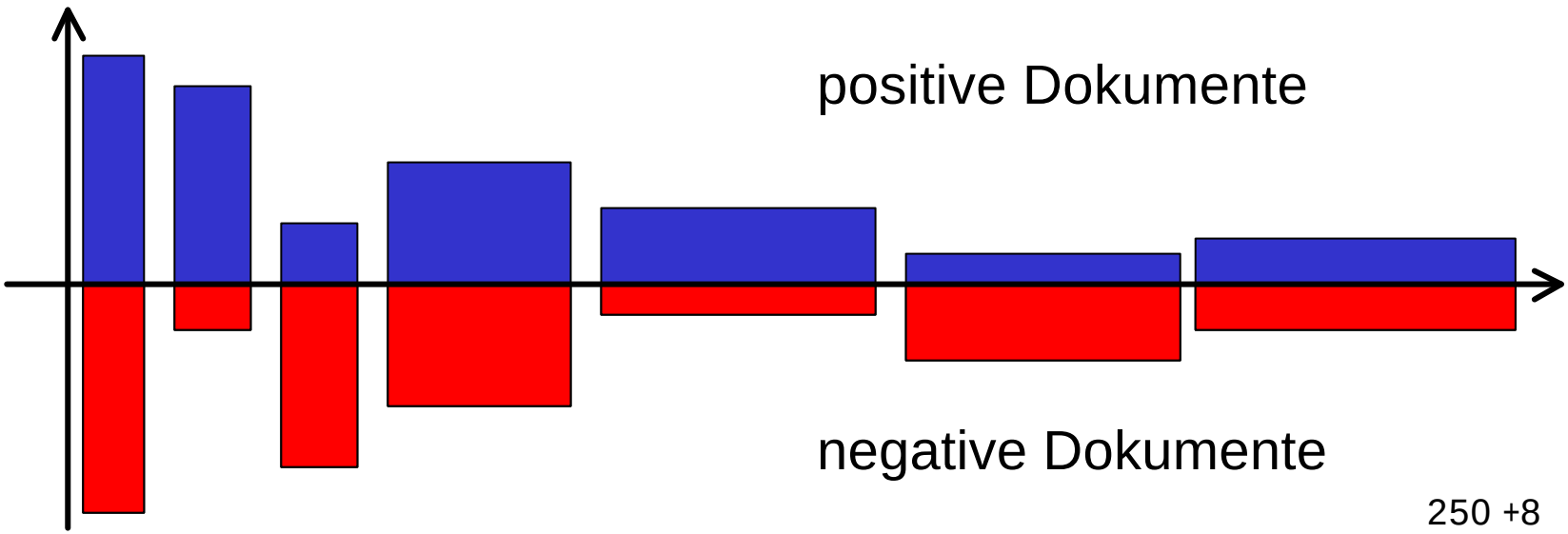
$$TFIDF(w_i, D) = \frac{TF(w_i, d) IDF(w_i, D)}{\sqrt{\sum_j \left[TF(w_j, d) IDF(w_j, D)\right]^2}}$$

Sind Texte nach Kategorien trennbar?

- Reuters, 10 Kategorien, TFIDF Repräsentation, C=50, 5 Kategorien ohne Fehler, bei den anderen durchschnittliche ξ-Werte-Summe 15 [2..33]
- WebKB, 4 Kategorien, TFIDF Repräsentation, C=50, keine Fehler.
- Ohsumed, 5 Kategorien, binäre Repräsentation, C=50, keine Fehler.
- Der Trainingsfehler zeigt an, dass Texte trotz der Hochdimensionalität, Heterogenität, Redundanz und Dünnbesetztheit gut trennbar sind.

TCat-Modell

- Typische Dimension: 10.000 – 100.000
- SVM lernt ohne Vorauswahl von Wörtern!
- **T**ext-**Cat**egorisierungs-Model (Joachims 2000)



TCAT

- TCat([p_1:n_1:f_1],..., [p_s:n_s:f_s]) mit s Mengen verschiedener Merkmale (Wörter) f_i, wobei in positiven Beispielen die Merkmale p mal und in negativen Beispielen die Merkmale n mal vorkommen.
- Wir nehmen sehr häufige, mittelhäufige und seltene Wörter in einem Text und betrachten, wie sie sich auf die positiven und negativen Beispiele verteilen.
- TCat ([20:20:100], //sehr häufig
  [4:1:200],[1:4:200],[5:5:600], //mittelhäufig
  [9:1:3000],[1:9:3000],[10:10:4000]) //selten
- Insgesamt 11 100 Wörter in 6 Gruppen.

Auswahl der Wortgruppen

- Odds ratio:
  a/b=c
  große c ergeben p
  kleine c ergeben n

Klasse	Merkmal	Ergebnis
POS	A	a
	notA	
NEG	A	b
	notA	

Hyperebene

$$w^* x + b = \sum_{i=1}^{11100} w_i x_i + b \qquad \text{b=0}$$

$$w_i = \begin{cases} +0{,}23 & \text{für die 200 mittelhäufigen pos Wörter} \\ -0{,}23 & \text{für die 200 mittelhäufigen neg Wörter} \\ +0{,}04 & \text{für die 3000 seltenen pos Wörter} \\ -0{,}04 & \text{für die 3000 seltenen neg Wörter} \\ 0 & \text{für die anderen Wörter} \end{cases}$$

TCat Lernbarkeit durch SVM

- TCat ist durch SVM lernbar! Joachims 2000 bzw. 2002
- Der erwartete Fehler einer SVM nach dem Training auf m Beispielen eines TCat Problems ist beschränkt durch:

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

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

wobei R^2 die maximale Euklidsche Länge eines Beispielvektors ist.

Bestimmung der Vektorlänge

- Nicht die Größe des Wörterbuchs ist wichtig, sondern wieviele verschiedene Wörter wirklich in einem Dokument vorkommen.
- Extremfälle:
  - Ein Dokument hat nur 1 Wort und das 1000 mal
    Dann ist die Euklidsche Länge des Vektors 1000
  - Ein Dokument hat 1000 Wörter und die kommen nur 1mal vor.
- Zipfsches Gesetz: Wörter werden in Häufigkeitsränge eingeteilt. Die r-häufigsten Wörter kommen 1/r so häufig vor wie die häufigsten. Also liegt die natürliche Sprache zwischen den Extremen. Der Vektor eines Textes hat also eine kürzere Euklidsche Länge als die Größe des Wörterbuchs oder die Länge des Textes vermuten lässt.

Beispiel: Intensivmedizin

$$f(x) = \left[\begin{pmatrix} 0.014 \\ 0.019 \\ -0.001 \\ -0.015 \\ -0.016 \\ 0.026 \\ 0.134 \\ -0.177 \\ \vdots \end{pmatrix} \begin{pmatrix} artsys = 174.00 \\ artdia = 86.00 \\ artmn = 121.00 \\ cvp = 8.00 \\ hr = 79.00 \\ papsys = 26.00 \\ papdia = 13.00 \\ papmn = 15.00 \\ \vdots \end{pmatrix} - 4.368 \right]$$

- Vitalzeichen von Intensivpatienten
- Dosis erhöhen oder nicht?
- Hohe Genauigkeit
- Verständlichkeit?

Schwierigkeit der Kategorie

- Die Fehlerabschätzung ergibt ein richtiges ranking der Schwierigkeitsgrade bei den untersuchten Textkategorien in TF-Repräsentation (Joachims 2000 bzw. 2002):
  - WebKB-Kategorie "course" hatte Trainingsfehler 0, Testfehler 4,4% und erwarteten Fehler nach TCat von 11,2%.
  - Reuters-Kategorie "earn" hatte Trainingsfehler 0, Testfehler 1,3% und erwarteten Fehler nach TCat von 1,5%.
  - Ohsumed-Kategorie "pathology" hatte Trainingsfehler 0, Testfehler 23,1% und erwarteten Fehler nach TCat von 94,5%.

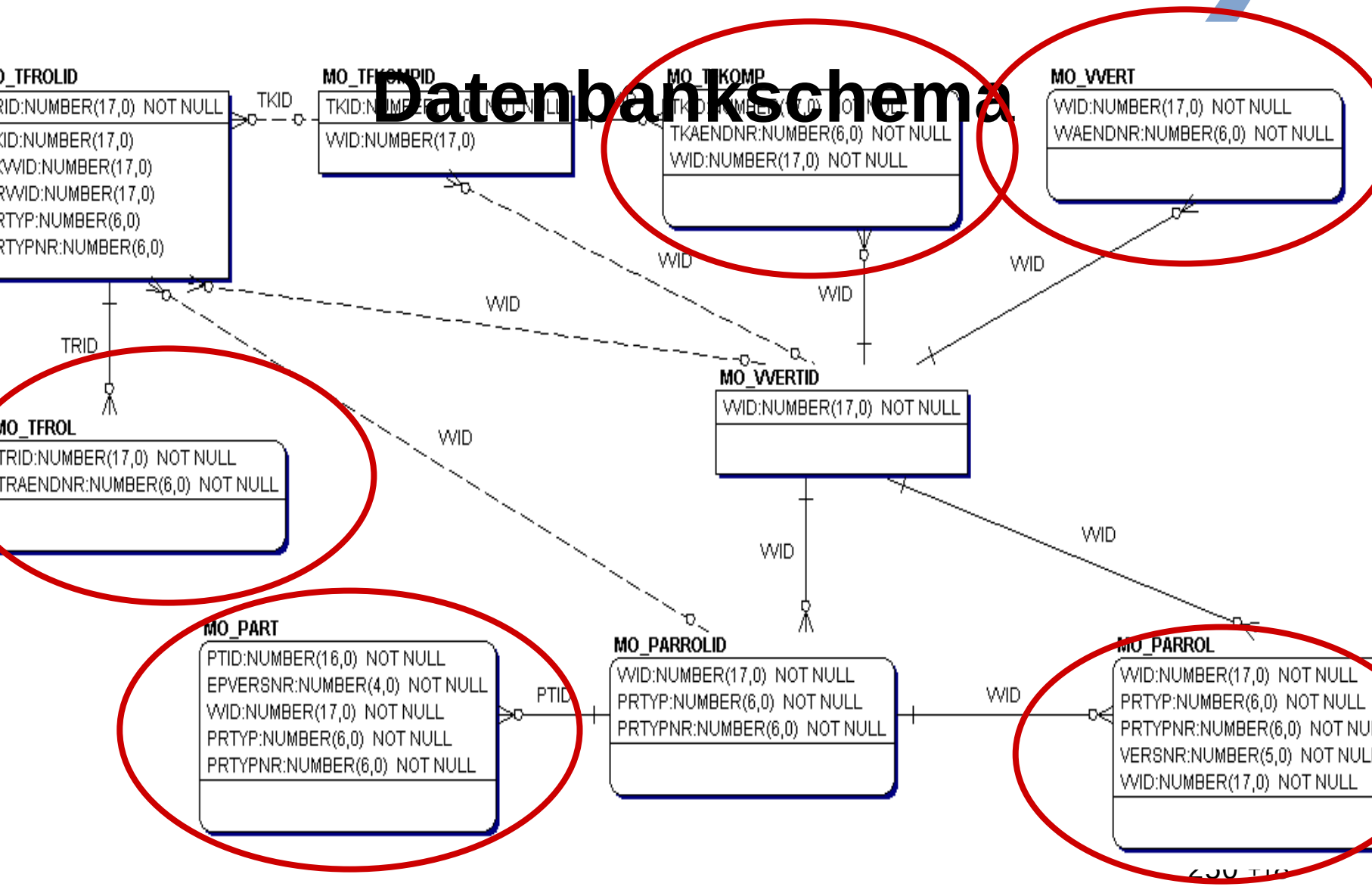
Begründete Faustregeln

- Repräsentation:
  - Wenn man nur TF verwendet, ist wichtig, dass im TCat Modell viele unterschiedliche (pi, ni) Wörter bei den hochfrequenten vorkommen (so in "course" [77:29:98]).
  - Wenn man TFIDF verwendet, wertet man die niedrigfrequenten Wörter auf (so in "pathology" – TFIDF führt zu 21,1% Testfehler).
- Das Verhältnis von pi und ni gibt ebenfalls die Schwierigkeit der Lernaufgabe nach TCat an: [9:6:500] >> [12:3:500] >> [15:0:500]
- Hohe Redundanz macht breiten Margin!
  Je größer p (ni) desto breiter der Margin.

Versicherungsdaten

- Auszug aus dem Data Warehouse einer Versicherungsgesellschaft in anonymisierter Form
- Oracle-Datenbank, 18 Tabellen und 15 Relationen
- Informationen zu Versicherungsverträgen und demographische Daten zu den Partnern
- 217 586 Versicherungsverträge und 163 745 Partner

Datenbankschema



Temporale Daten

- Tupelzeitstempelung
- Eindeutige Änderungsnummer
- Gültigkeitsdauer (valid time)
- In der Vertragstabelle zusätzlich Abwicklungszeit (transaction time) und Änderungsgrund

Auszug aus der Versicherungstabelle

VVID	VVAENDR	VVWLVO	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	
...	...	...	...	...	...	250 +20

Aufgabenstellung

- Vorhersage des Rückkaufs von Versicherungen
- Rückkauf :
  - Vorzeitige Beendigung einer Versicherung wegen Rücktritt oder Kündigung
  - dem Versicherungsnehmer ist der aktuelle Zeitwert auszuzahlen
- Rückkaufe sind Verlustgeschäft für die Versicherungsgesellschaft

Keine Berücksichtigung der Zeit

- Hypothese: Kundendaten geben Aufschluss über einen möglichen Rückkauf
- SVM, Entscheidungsbaumlerner und Apriori
- Bestes Ergebnis:
  - Precision: 57%
  - Recall: 80%

Berücksichtigung der Zeit

- Häufige Änderungsmuster
- Regeln, die Rückkauf vorhersagen können
- Algorithmus, der auf dem Ansatz von Höppner basiert
- Erzeugung von Intervallen aus den Gültigkeitszeitmarken der Tupel
- Allens zeitliche Intervalllogik zur Formulierung von Relationen zwischen den erzeugten Intervallen
- Regeln liefern keine Hinweise, warum ein Rückkauf erfolgt.

Fazit

- Die Daten enthalten keine relevanten Informationen, um Rückkauf vorhersagen zu können
- Die Repräsentation der Daten wurde nicht geeignet gewählt
- Neue Vermutung:

  Häufige Änderungen eines Vertrages sind Ausdruck von Unzufriedenheit des Kunden mit dem Vertrag
- Ansatz: Häufigkeitsbasierte Repräsentation

- Information Retrieval **TFIDF**
- Textkategorisierung: Gewichtung der Wörter
- Die *Termfrequenz (term frequency)* tf(w,d) gibt an, wie oft das Wort w im Dokument d auftritt
- Die *Dokumentfrequenz (document frequency)* df(w) ist die Anzahl der Dokumente, in denen das Wort w mindestens einmal auftritt
- *Inverse Document Frequency*:

$$idf(w) = \log\left(\frac{|D|}{df(w)}\right)$$

- *TFIDF-Gewicht* eines Wortes w in einem Dokument d:

$$tfidf(w,d) = tf(w,d) * idf(w)$$

Merkmalsgenerierung mit Hilfe von TFIDF

- Termfrequenz beschreibt, wie oft ein bestimmtes Attribut in einem Vertrag geändert wurde

$$tf(a_i, c_j) = \left| \{ x \in \textbf{Zeitpunkte} \mid a_i \textbf{ wurde geändert} \} \right|$$

- Die Dokumentfrequenz entspricht der Anzahl der Verträge, in denen das Attribut geändert wurde

$$df(a_i) = \left| \{ c_j \in C \mid a_i \textbf{ wurde geändert} \} \right|$$

- TFIDF Merkmale

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

Erzeugung der TFIDF Merkmale

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

Lernverfahren und Ergebnisse

- Training einer SVM
- 10-fache Kreuzvalidierung
- Ergebnis
  - Accuracy: 99,4%
  - Precision: 94,9%
  - Recall: 98,2%
- Sind die guten Ergebnisse Zufall?

TCat-Konzepte

- Das TCat-Konzept

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

beschreibt eine binäre Klassifikationsaufgaben mit s disjunkten Mengen von Merkmalen. Die i-te Menge enthält f_i Merkmale. Jedes positive Beispiel enthält p_i Merkmale aus der jeweiligen Menge, und jedes negative Beispiele enthält n_i Merkmale. Das gleiche Merkmal kann mehrmals in einem Dokument vorkommen.

Häufigkeit der Attribute

- Einteilung der Attribute in hoch (high frequency)-, mittel (medium frequency)- und niedrigfrequente (low frequency) Attribute
- Einteilung in positive und negative Indikatoren anhand der OddsRatio-Werte

	high frequency	medium frequency	low frequency	Rest
Positive Indikatoren	2	3	19	39
Negative Indikatoren	3	4	64	250 +30

Zusammensetzung eines Vertrages

- Von Details eines einzelnen Vertrages abstrahieren
- Durchschnittlicher Vertrag hat 8 Merkmale

	high frequency		medium frequency		low frequency		
	2 pos.	3 neg.	3 pos.	4 neg.	19 pos.	64 neg.	39 rest
Positiver Vertrag	25%	12,5%	37,5%	0%	12,5%	0%	12,5%
Negativer Vertrag	0%	50%	12,5%	12,5%	0%	12,5%	12,5%

Modellierung als Tcat-Konzept

- 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
)

Lernbarkeit von TCat-Konzepten

- Schranke des erwarteten Generalisierungsfehles einer Support Vector Maschine nach Joachims

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

mit

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

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

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

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

Erwarteter Generalisierungsfehler

- Nach der Mandelbrot Verteilung ist $R^2 \leq 37$
- Schranke für den erwarteten Fehler:

$$e\left(Err^n\left(h_{SVM}\right)\right) \leq \frac{22}{n+1}$$

- Nach einem Training auf 1000 Beispielen beträgt der erwartete Fehler weniger als 2,2%

Results (F-measure)

Learner	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