



Mit KI gegen SPAM

Proseminar Künstliche Intelligenz

SS 2006

Florian Laib



Ausblick

- Was ist SPAM? Warum SPAM-Filter?
- Naive Bayes-Verfahren
- Fallbasiertes Schließen
- Fallbasierte Filter TiMBL
- Vergleich der beiden Verfahren
- Zusammenfassung und Bewertung



Was ist SPAM?

- Definition:
unverlangt zugesandte **Massen-** bzw. **Werbe-E-Mail**
- Beispiele: Kettenbriefe, Hoaxes (Scherzmeldungen), Phishing-Mails, Werbe-Mails, etc.
- Top Werbe-Mails: Arzneimittel (41.4%), Kreditangebote (11.1%), Pornographische Inhalte (9.5%)
- Woher kommt SPAM? 1.Platz USA (24.5%), 2.Platz China (22.3%) 3.Platz Südkorea (9,7%)
- SPAM lohnt sich für Spammer bereits, wenn von 5 Millionen SPAMs nur 5 Menschen ein Produkt kaufen



Warum SPAM-Filter?

- SPAM Anteil liegt heute bei ca. 60% - 80%
- Verschwendung von Bandbreite und längere Downloadzeiten
- Überlastung von Servern, da zuerst SPAMs „abgearbeitet“ werden. Folge: eigener E-Mail-Versand dauert länger
- Volle bzw. Überfüllte Mailboxen
- Folge: möglicher Verlust von erwünschten Nachrichten
- Manuelles Filtern dieser Nachrichten nimmt viel Zeit in Anspruch
- Enorme Kosten für Unternehmen und Privatanwender
- Schätzung: USA 10 Mrd. \$ pro Jahr, Europa 3 Mrd. \$ (2004)



Warum SPAM-Filter?

- Verwendung von Filtern senkt Kosten und spart Zeit
- Filter werden immer besser
- Spammer suchen neue Wege um Filter zu umgehen
- z.B. Einfügen „guter“ Wörter, Verändern von Wörtern
- Aus Viagra wird dann V*i*a*g*r*a oder V i a g r a
- SPAMs verändern sich ständig-> Filter müssen sich anpassen
- Manuelle Anpassung kostet wieder Zeit und Geld
- Lösung: lernende Filter
- Lernen: Gesetzmäßigkeiten anhand von Beispielen erkennen



False Positive/Negative

- Zwei mögliche Fehler beim Filtern von E-Mails
- Nachricht irrtümlicherweise als SPAM klassifizieren
- Bezeichnet als **False Positive**
- Nachricht irrtümlicherweise als erwünschte Nachricht klassifizieren
- Bezeichnet als **False Negative**



Überblick

- Was ist SPAM? Warum SPAM-Filter?
- **Naive Bayes-Verfahren**
- Fallbasiertes Schließen
- Fallbasierte Filter TiMBL
- Vergleich der beiden Verfahren
- Zusammenfassung und Bewertung

Naive Bayes-Verfahren

- Bisher bevorzugtes Verfahren
- Basiert auf Bayes-Theorem (Thomas Bayes)
- Als Formel:

$$P(h | D) = \frac{P(D | h)P(h)}{P(D)}$$

Daten bzw. E-Mail Nachricht

Hypothese, also SPAM oder Nicht-SPAM

- Benutzt Wahrscheinlichkeiten um neue Nachricht zu klassifizieren



Naive Bayes-Verfahren

- Berechnung der Wahrscheinlichkeiten $P(D|h)$, $P(h)$ und $P(D)$ relativ einfach
- Man braucht nur manuell klassifizierte SPAM und Nicht-SPAM Nachrichten
- Berechnet aus Wahrscheinlichkeit das Wörter in SPAM auftreten die „umgedrehte“ Wahrscheinlichkeit, dass eine Nachricht SPAM ist wenn bestimmte Wörter darin auftreten
- Klassifizierung als SPAM, falls bestimmte Schwelle überschritten wird (z.B. SPAM, falls $P(h|D) > 95\%$)
- Benutzt von z.B. Mozilla Thunderbird und Microsoft Outlook
- Grund: funktioniert, liefert gute Ergebnisse und ist einfach zu Implementieren



Naive Bayes-Verfahren

- Heißt „naiv“, da angenommen wird, dass alle Wörter unabhängig voneinander auftreten
- z.B. „Grüßen“ oft zusammen mit „freundlichen“ auftritt
- Filter lernt, indem durch neue Nachrichten alle Wahrscheinlichkeiten neu berechnet und angepasst werden
- Es werden nicht die Nachrichten sondern nur die Wahrscheinlichkeiten gespeichert

- Problem: Spammer fügen „gute“ Wörter in SPAM ein
- Kann dazu führen, dass Nachricht nicht als SPAM erkannt wird



Naive Bayes - Einfaches Beispiel 1/2

- Annahme: Hälfte meiner E-Mails sind SPAM
 $P(\text{SPAM}) = 0.5$ und $P(\text{Nicht-SPAM}) = 0.5$
- Sei Wahrscheinlichkeit für das Wort „Viagra“ in SPAM 10%
 $P(\text{„Viagra“} \mid \text{SPAM}) = 0.1$
- In Nicht-SPAM Nachrichten kommt es mit 1% vor
 $P(\text{„Viagra“} \mid \text{Nicht-SPAM}) = 0.01$
- $P(\text{„Viagra“}) = P(\text{„Viagra“} \mid \text{SPAM}) * P(\text{SPAM}) + P(\text{„Viagra“} \mid \text{Nicht-SPAM}) * P(\text{Nicht-SPAM}) = 0.1 * 0.5 + 0.01 * 0.5 = 0.055$ (=5.5%)



Naive Bayes - Einfaches Beispiel 2/2

- Bayes-Verfahren liefert Antwort auf die Frage:
„Wie groß ist die Wahrscheinlichkeit das eine Nachricht SPAM ist, falls „Viagra“ darin vorkommt?“
- $P(\text{SPAM} | \text{„Viagra“}) = (P(\text{„Viagra“} | \text{SPAM}) * P(\text{SPAM})) / P(\text{„Viagra“})$
 $P(\text{SPAM} | \text{„Viagra“}) = (0.1 * 0.5) / 0.055 = 0.9091 (=90.91\%)$
- Wäre $P(\text{„Viagra“} | \text{Nicht-SPAM}) = 0 (=0\%, \text{ d.h. kommt nur in SPAM vor})$, ergibt sich:
- $P(\text{SPAM} | \text{„Viagra“}) = (P(\text{„Viagra“} | \text{SPAM}) * P(\text{SPAM})) / P(\text{„Viagra“})$
 $P(\text{SPAM} | \text{„Viagra“}) = (0.1 * 0.5) / 0.05 = 1 (=100\%)$



Zusammenfassung

- Bevorzugtes Verfahren
- Bayes Theorem
- Einfache Berechnung der Wahrscheinlichkeiten
- Funktioniert mit guten Ergebnissen
- Lernen durch neue Nachrichten und Anpassen der Wahrscheinlichkeiten



Überblick

- Was ist SPAM? Warum SPAM-Filter?
- Naive Bayes-Verfahren
- **Fallbasiertes Schließen**
- Fallbasierte Filter TiMBL
- Vergleich der beiden Verfahren
- Zusammenfassung und Bewertung



Fallbasiertes Schließen

- Neuer, anderer Ansatz
- Idee: Lösen eines neuen Problems durch das Erinnern an eine frühere ähnliche Situation und die Verwendung der Informationen und des Wissens dieser Situation.
- Fallbasis (case base), besteht aus manuell klassifizierten SPAM und Nicht-SPAM Nachrichten
- Klassifiziert neue Nachricht durch suchen und analysieren der k ähnlichsten Nachrichten der Fallbasis
- Neue Nachricht wird der Klasse zugeordnet, der die Mehrheit der ähnlichsten Nachrichten zugehört



Fallbasiertes Schließen

- Alle Nachrichten werden gespeichert und bei der Klassifizierung berücksichtigt
- Lernen erfolgt einfach dadurch, dass neue Nachrichten der Fallbasis hinzugefügt werden
- Lernphase ist deshalb einfacher als beim NB-Verfahren
- Neue Nachrichten stehen sofort der Klassifizierung zur Verfügung
- Klassifizierungsphase selber ist im Gegensatz zum NB-Verfahren aber zeitaufwändiger
- Anstatt „nur“ eine Formel zu berechnen (NB) müssen die k ähnlichsten Nachrichten gefunden und analysiert werden



Zusammenfassung

- Neuer Ansatz
- Fallbasis aus SPAM und Nicht-SPAM
- Klassifizierung durch Ähnlichkeit zu „alten“ Nachrichten
- Alle Nachrichten werden gespeichert und berücksichtigt
- Lernen: Neue Nachrichten der Fallbasis hinzufügen



Überblick

- Was ist SPAM? Warum SPAM-Filter?
- Naive Bayes-Verfahren
- Fallbasiertes Schließen
- **Fallbasierte Filter TiMBL**
- Vergleich der beiden Verfahren
- Zusammenfassung und Bewertung



Fallbasierter Filter TiMBL

- TiMBL steht für Tilburg Memory Based Learner
- Problem: Woher bekommt man SPAM und Nicht-SPAM Nachrichten, die auch veröffentlicht werden können?
- SPAM Nachrichten kein Problem, da schon „öffentlich“
- Problem bei Nicht-SPAM Nachrichten wegen Privatsphäre
- Lösung: Mailing Listen
- Fallbasis: 2893 E-Mails einer Liste über Sprachwissenschaften
- 2412 Nicht-SPAM und 481 SPAM Nachrichten
- Entspricht ca. 16% SPAM (heute bis zu 80% SPAM)



Darstellung der Nachrichten

- E-Mail Nachricht wird als Vektor dargestellt

$$\vec{x} = (x_1, x_2, \dots, x_n)$$

- $X_i = 1$, falls das Wort in der Nachricht ist, $X_i = 0$, falls nicht
- Anstatt binärer Darstellung auch numerische Darstellung möglich
- Aufbereitung der Wörter:
 - Selten vorkommende Wörter wurden weggelassen
 - Wörter auf Stammformen bringen, z.B. aus „was“ wird „be“
 - Reduzieren der Anzahl von Wörtern auf die wichtigsten bzw. aussagekräftigsten für die Klassifizierung

K–Nächste-Nachbarn 1/2

- K–Nächste-Nachbarn Verfahren zur Bestimmung der k ähnlichsten Nachrichten
- TiMBL verwendet sog. overlap Verfahren

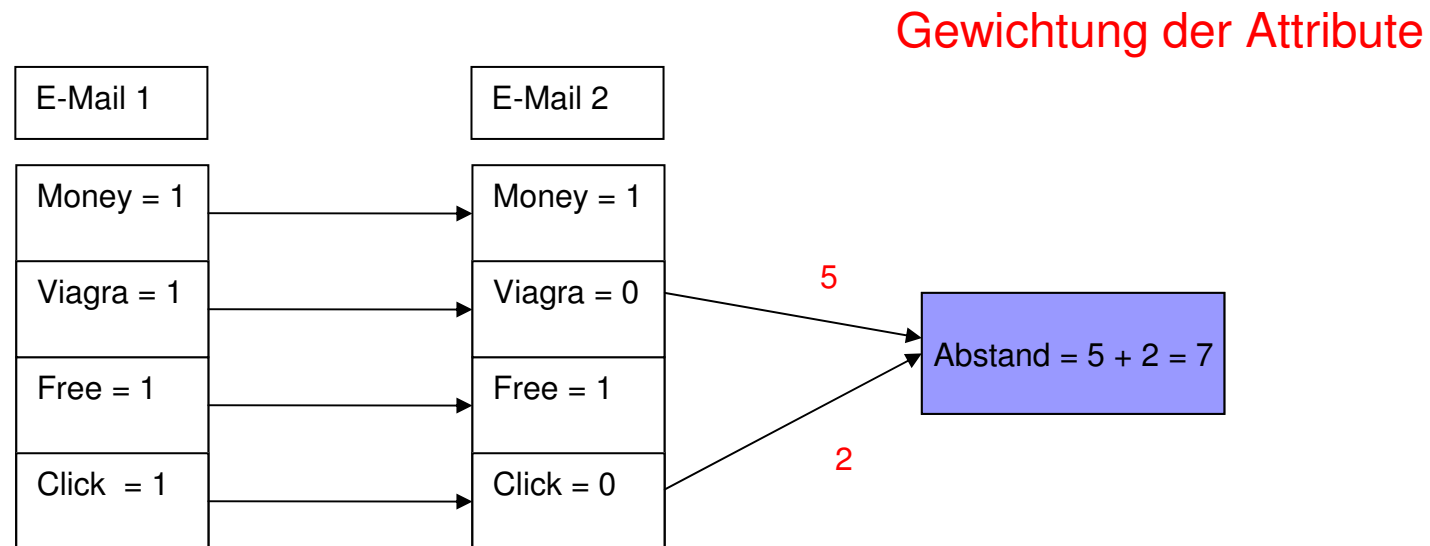
$$d(\vec{x}_i, \vec{x}_j) \equiv \sum_{r=1}^n \delta(x_{ir}, x_{jr})$$

wobei

$$\delta(x, y) \equiv \begin{cases} 0, & x = y \\ 1, & \text{sonst} \end{cases}$$

K-Nächste-Nachbarn 2/2

- Einfaches Beispiel:

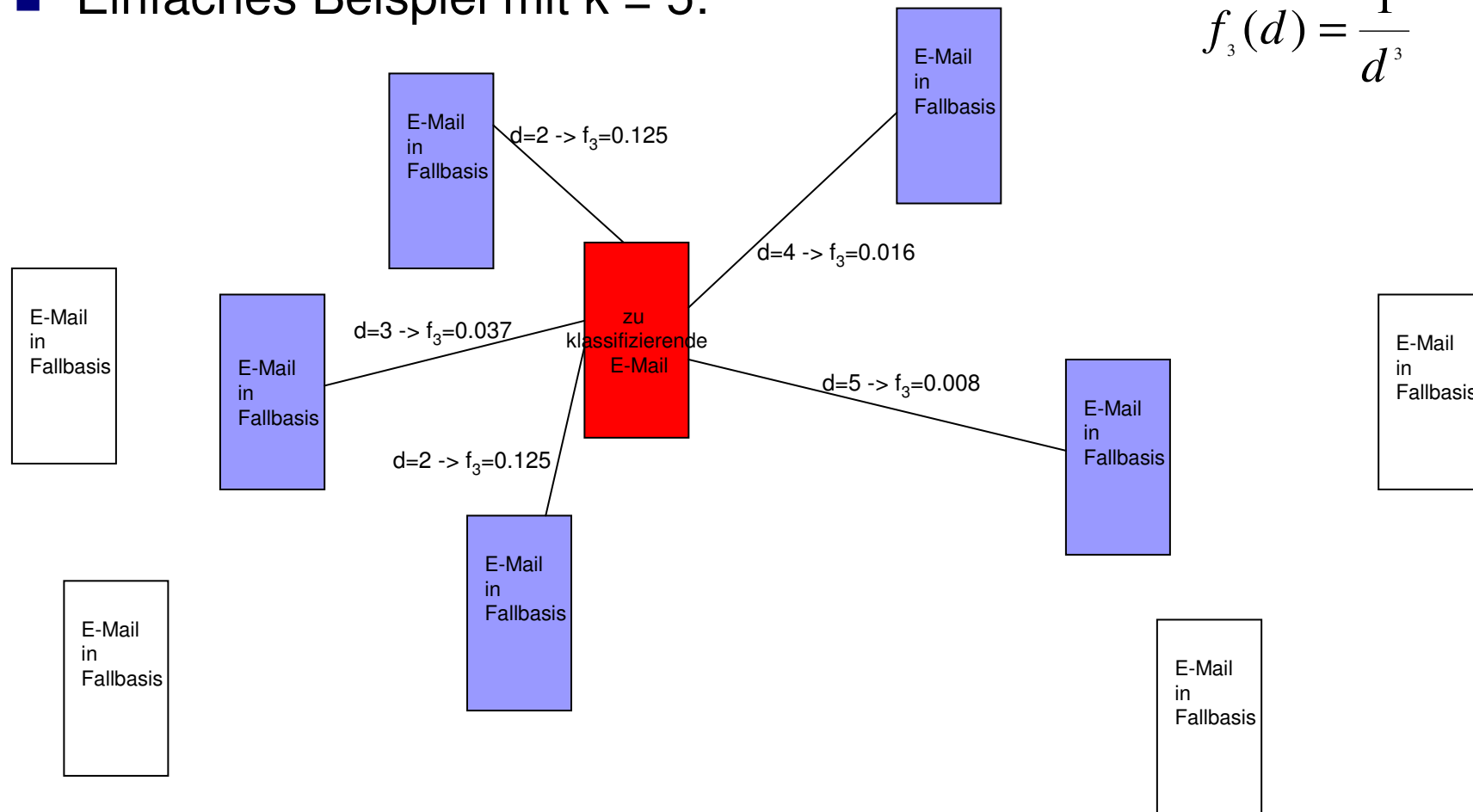


Distance weighting

Gewichtung der Abstände

- Einfaches Beispiel mit $k = 5$:

$$f_3(d) = \frac{1}{d^3}$$





Cost-sensitive classification

- False Positive schlimmere Folgen als False Negative
- Kosten entsprechen Konsequenzen für Anwender bzw. Aufwand die Folgen/Fehler rückgängig zu machen
- False Negative mit Kosten = 1 und False Positive = λ
- Nachricht als SPAM klassifiziert, falls Kosten um sie als Nicht-SPAM zu klassifizieren größer als Kosten um als SPAM zu klassifizieren
- D.h. Nachricht ist SPAM, falls eine sog. Klassifizierungsschwelle t überschritten wird mit $t = \lambda/(1 + \lambda)$
- Falls Nachricht(en) mit Abstand $d = 0$ gefunden, Klassifizierung nach Mehrheit dieser Nachrichten



Überblick

- Was ist SPAM? Warum SPAM-Filter?
- Naive Bayes-Verfahren
- Fallbasiertes Schließen
- Fallbasierte Filter TiMBL
- **Vergleich der beiden Verfahren**
- Zusammenfassung und Bewertung



Vergleich der beiden Verfahren

- **Recall:** Anteil der SPAM Nachrichten die richtigerweise als SPAM klassifiziert werden
- **Precision:** Anteil der SPAM Nachrichten an den insgesamt blockierten Nachrichten
- **TCR (total cost ratio):** Wert der die Kosten der Klassifizierung berücksichtigt. Je höher der Wert desto besser der Filter. Ist der Wert kleiner als 1 wäre es besser den Filter nicht zu benutzen.



Vergleich der beiden Verfahren

- Beste Resultate des TiMBL Filters:

λ	Recall (%)	Precision (%)	TCR
1	88.60	97.39	7.18
9	81.93	98.79	3.64
999	59.91	100	2.49

- Beste Resultate für einen naiven Bayes-Filter:

λ	Recall (%)	Precision (%)	TCR
1	82.35	99.02	5.41
9	77.57	99.45	3.82
999	63.67	100	2.86



Vergleich der beiden Verfahren

- Fallbasierte Filter liefert mindestens gleich gute Ergebnisse wie ein naiver Bayes Filter
- NB-Filter speichert nur die Wahrscheinlichkeiten
- Geringerer Speicherbedarf und Privatsphäre gewährleistet
- Aber sehr schlecht nachzuvollziehen wie die Werte zustande kommen
- NB-Filter schneller zur Klassifizierungszeit



Vergleich der beiden Verfahren

- Fallbasierter Filter speichert komplette Nachrichten
- Größerer Speicherbedarf
- Erklärbarkeit warum eine Nachricht als SPAM klassifiziert wurde
- Man kann sich z.B. die k Nachbarn anschauen, etc.
- Lernphase bei Fallbasierten Filter schneller und leichter
- Zu „alte“ Nachrichten können einfach weggelassen werden



Überblick

- Was ist SPAM? Warum SPAM-Filter?
- Naive Bayes-Verfahren
- Fallbasiertes Schließen
- Fallbasierte Filter TiMBL
- Vergleich der beiden Verfahren
- **Zusammenfassung und Bewertung**



Zusammenfassung und Bewertung

- Performance bzw. Ergebnisse sehr schwer zu vergleichen
- Keine einheitlichen Trainingsdaten
- Nachrichten von Mailing Listen nicht mit Nachrichten von Privatanwendern vergleichbar
- Quellen sind schon mehrere Jahre alt
- Geringe SPAM Anteil von nur 16% heute bis zu 80%
- Fehlen von neuen SPAM (-Techniken)
- Fehlende Berücksichtigung von Bildern, Anhänge, etc.



Zusammenfassung und Bewertung

- Naive Bayes-Verfahren:
 - Bayes Formel zur Klassifizierung
 - Wahrscheinlichkeiten leicht zu berechnen
 - Speichert nur Werte
 - Liefert gute Ergebnisse

- Fallbasiertes Schließen:
 - Fallbasis aus SPAM und Nicht-SPAM Nachrichten
 - Klassifizierung durch Ähnlichkeit zu „alten“ Nachrichten
 - Nachrichten werden gespeichert
 - Leichte und schnelle Lernphase
 - Besonders gute Ergebnisse bei Aktualisieren der Fallbasis