



**Fachgebiet Wissensmanagement**

**Prof. Dr. Tobias Scheffer**

Studienarbeit

**Nächste-Nachbarn-Klassifikatoren  
zur automatisierten E-Mail-Beantwortung**

Eingereicht von: Thomas Posner  
posner@informatik.hu-berlin.de  
Matr.-Nr. 158077

Betreuer: Prof. Dr. Tobias Scheffer, Steffen Bickel

19. Dezember 2004

## Zusammenfassung

Mit dieser Arbeit werden drei Nächste-Nachbarn-Methoden zur automatischen Beantwortung von E-Mails entwickelt und hinsichtlich ihrer Leistungsfähigkeit untersucht. Der Datenbestand aus zwei Fallstudien dient dabei als Trainings- und Bewertungsgrundlage. Wir werden die Problemstellung beschreiben, die Nächste-Nachbarn-Verfahren (NN-Verfahren) erläutern sowie Methoden aus anderen Arbeiten vorstellen. Danach evaluieren wir die Testdatensätze, wenden die programmierten Methoden an und werten sie aus. Im Besonderen vergleichen wir die NN-Verfahren mit zwei anderen Ansätzen. Im Ergebnis konnten wir für die automatische Beantwortung von E-Mails die NN-Verfahren als untere Leistungsgrenze bestätigen.

## Inhaltsverzeichnis

<b>1. EINLEITUNG</b> .....	<b>1</b>
<b>2. PROBLEMSTELLUNG</b> .....	<b>2</b>
<b>3. NN-KLASSIFIKATOREN</b> .....	<b>4</b>
3.1. ALLGEMEINE FUNKTIONSWEISE.....	4
3.2. DAS VEKTORRAUMMODELL.....	6
3.3. ANPASSUNGEN AN DIE E-MAIL-BEANTWORTUNG .....	8
3.3.1. <i>NN-Klassifikation mit den Fragen</i> .....	9
3.3.2. <i>NN-Klassifikation mit den Fragen und Antworten</i> .....	10
3.3.3. <i>NN-Klassifikation mit den Fragen und Antworten (Ähnlichkeitsmittelung der Fragen)</i> .....	11
<b>4. REFERENZMETHODEN ZUR AUTOMATISCHEN E-MAIL-BEANTWORTUNG</b> .....	<b>11</b>
4.1. ÜBERWACHTE KLASSEIFIKATION.....	11
4.2. LERNEN AUS FRAGE-ANTWORT-PAAREN.....	12
<b>5. DATENSÄTZE ZUR EVALUIERUNG</b> .....	<b>13</b>
5.1. ONLINE-VERSANDHÄNDLER .....	14
5.1.1. <i>Verifikation der Tauglichkeit</i> .....	14
5.2. EDV-HOTLINE IM KRANKENHAUS.....	15
5.2.1. <i>Verifikation der Tauglichkeit</i> .....	16
<b>6. AUSWERTUNG</b> .....	<b>17</b>
6.1. ANWENDUNG DER NN-KLASSIFIKATOREN .....	17
6.1.1. <i>NN-Klassifikation mit den Fragen</i> .....	18
6.1.2. <i>NN-Klassifikation mit den Fragen und Antworten</i> .....	19
6.1.3. <i>NN-Klassifikation mit den Fragen und Antworten (Ähnlichkeitsmittelung der Fragen)</i> .....	20
6.2. VERGLEICH ÄHNLICHER METHODEN.....	21
<b>7. SCHLUSSFOLGERUNGEN</b> .....	<b>23</b>
<b>LITERATUR- UND QUELLENVERZEICHNIS</b> .....	<b>25</b>

## Abbildungsverzeichnis

Abbildung 1: Beispiel kNN-Methode [Mitchel, 1997].....	5
Abbildung 2: Voronoi Diagramm für $k=1$ [Mitchel, 1997].....	6
Abbildung 3: Unterschiedliche Ergebnisse der kNN-Methode bei Variation von $k$ [iLink01] .....	6
Abbildung 4: SV-Klassifikation der Daten des Online-Versandhändlers .....	15
Abbildung 5: Klasseneinteilung der EDV-Hotline-E-Mails.....	16
Abbildung 6: SV-Klassifikation der EDV-Hotline-Daten.....	17
Abbildung 7: kNN-Klassifikation mit den Fragen .....	18
Abbildung 8: kNN-Klassifikation mit den Fragen und Antworten.....	19
Abbildung 9: kNN-Klassifikation mit den Fragen und Antworten (Ähnlichkeitsmittelung Fragen) .....	21
Abbildung 10: Vergleich verschiedener Methoden zur E-Mail-Beantwortung.....	22
Abbildung 11: Unterschiedliche Mittelpunkte bei Anwendung der NN-Klassifikation mit den Fragen und Antworten .....	23

## 1. Einleitung

Ein Großteil der heutigen Kommunikation wird per E-Mail abgewickelt. In Firmen verwenden Mitarbeiter einen bedeutenden Anteil ihrer Arbeitszeit für das Lesen und Beantworten von E-Mails. Domänen mit häufig identischen E-Mail-Anfragen würden durch die Automatisierung des Beantwortungsprozesses viel Zeit sparen und dem Unternehmen einen Kostenvorteil gegenüber der Konkurrenz verschaffen.

Verschiedene ähnliche Probleme wurden schon in anderen Arbeiten untersucht. Die Erkennung von Spam-E-Mails wurde ausführlich in [Cohen, 1996], [Sahamie et. al, 1998] und [Ducker et. al, 1999] erforscht. Ebenso konnte das Sortieren von E-Mails in Ordner sowie die Priorisierung und Weiterleitung von E-Mails untersucht werden ([Boone, 1998]).

Mit dieser Arbeit wird die Leistungsfähigkeit der Nächste-Nachbarn-Klassifikatoren (NN-Klassifikatoren) untersucht, um E-Mails automatisiert zu beantworten. Das bedeutet hier, für Klassen von E-Mail-Anfragen vorab Antworten zu formulieren und diese einer neuen E-Mail zuzuweisen. Die Idee wurde mit der Arbeit [Kockelkorn/Lüneburg/Scheffer, 2003] implementiert. Die Leistungsfähigkeit ähnlicher Methoden konnten [Bickel/Scheffer, 2004] untersuchen.

Zum Training und zur Bewertung stehen die Testdatensätze von den Kundenservicecentern eines Rechenzentrums sowie von einem großen Onlinehändler zur Verfügung. Ein Testdatensatz besteht aus den gesendeten E-Mails eines Mitarbeiters. Da in Kundenservicecentern häufig E-Mails mit ähnlichen Anfragen eintreffen, existieren zwischen diesen genug Ähnlichkeiten um maschinelle Algorithmen für die Beantwortung anzuwenden.

Die Arbeit beschreibt zuerst die Problemstellung und die Grundlagen der angewendeten NN-Klassifikatoren. Sie beschreibt ausführlich die Anpassungen der Nächste-Nachbarn-Methode (NN-Methode) an die E-Mail-Beantwortung und geht kurz auf verwendete

Methoden in ähnlichen Arbeiten ein. Darauf folgend werden die Datensätze samt ihren Eigenschaften vorgestellt und auf ihre Eignung zur automatischen Beantwortung von E-Mails überprüft. Zudem werden die mit dieser Arbeit programmierten NN-Methoden hinsichtlich der Leistungsfähigkeit an den Testdaten ausgewertet. Abschließend vergleichen wir die Leistungsdaten der NN-Methoden mit Ansätzen aus ähnlichen Arbeiten und fassen unsere Erkenntnisse zusammen.

## 2. Problemstellung

Die Daten bestehen aus eingehenden E-Mails  $x \in X$  und ausgehenden E-Mails  $y \in Y$ . Das Beantwortungsproblem kann somit durch die Verteilung  $p(x,y)$  und einer Akzeptanzfunktion  $akzept: (x, y) \mapsto \{0,1\}$  beschrieben werden. Die Akzeptanzfunktion entscheidet, ob eine E-Mail akzeptabel beantwortet wurde. Weder die Verteilung noch die Akzeptanzfunktion sind bekannt.  $X$  und  $Y$  sind die Menge aller möglichen Zeichenketten und dementsprechend die Menge aller möglichen E-Mails. Die Abbildung der Fragen auf die Antworten ist das zu lösende Lernproblem.

Die Trainingsdaten können nun als eine Sequenz der Form  $M = \{(x_1, y_1), \dots, (x_m, y_m)\}$  dargestellt werden, wobei die Verteilung durch  $p(x,y)$  bestimmt wird. Für jedes Trainingspaar  $(x, y)$  aus den Testdaten liefert die Akzeptanzfunktion den Wert 1.

Wenn eine neue E-Mail beantwortet wird, dann ist  $akzept(x, f(x)) = 1$  der optimale und  $akzept(x, f(x)) = 0$  der nicht wünschenswerte Fall.

Um die Genauigkeit (Precision) der Beantwortung zu messen, stehen die akzeptabel beantworteten E-Mails im Verhältnis zu allen beantworteten E-Mails.

$$\text{Precision} = \frac{\text{akzeptabel beantwortete E - Mails}}{\text{alle beantworteten E - Mails}} \quad (2.1)$$

Der Wiedererkennungswert (Recall) setzt die akzeptabel beantworteten E-Mails in das Verhältnis zu allen E-Mails.

$$\text{Recall} = \frac{\text{akzeptabel beantwortete E - Mails}}{\text{alle E - Mails}} \quad (2.2)$$

Diese Definitionsanpassung von Precision und Recall an die E-Mail-Domäne ist abgeleitet aus [Manning/Schütze, 1999] und geht einher mit der Definition von [Bickel/Scheffer, 2004].

Die E-Mail-Beantwortungsmethoden nutzen einen Grenzwert  $\Theta$ , der die Überzeugung der Methode in Bezug auf die Antwort widerspiegelt.

Durch kleine Grenzwerte ist der Recall hoch und die Precision klein. Mit einem steigenden Grenzwert verringert sich die Anzahl an falschen Antworten und die Precision steigt an. Gleichzeitig verringert sich die Gesamtzahl der akzeptabel beantworteten E-Mails, wodurch der Recall reduziert wird.

Mit dieser schrittweisen Erhöhung des Grenzwertes  $\Theta$  entstehen Precision-Recall-Kurven (PR-Kurven), die eine Möglichkeit darstellen, um die Leistungsfähigkeit eines Lernverfahrens einzuschätzen [Manning/Schütze, 1999].

Die akzept-Funktion soll die Entscheidung eines Menschen nachstellen. Um nun auch diese Funktion während des Trainings automatisiert auszuführen, werden die Testdaten vorab manuell in Klassen  $S = \{S_1, \dots, S_N\}$  eingeteilt, so dass jede Klasse semantisch äquivalente E-Mails enthält. Die Klasse  $S_N$  (sonstige E-Mails) enthält alle E-Mails, für die keine hinreichend große Klasse gebildet werden konnte. Die akzept-Funktion greift nun in ihrer automatischen Bewertung auf diese Klassen zu. Für ein Trainingsbeispiel  $(x_j, y_j) \in M$  gibt die Funktion  $\text{akzept}(x_j, f(x_j)) = 1$  zurück, falls  $f(x_j)$  und  $y_j$  in der selben Klasse  $S_i$  liegen und  $S_i \neq S_N$  gilt. Andernfalls ist  $\text{akzept}(x_j, f(x_j)) = 0$ .

### 3. NN-Klassifikatoren

Die NN-Verfahren werden auch als K-Nächste-Nachbarn-Verfahren bezeichnet, da zur Klassifikation eines neuen Beispiels K Nachbarn hinzugezogen werden.

Zuerst wird in Anlehnung an [Mitchel, 1997] im Abschnitt 3.1. die allgemeine Funktionsweise der NN-Klassifikatoren erläutert. Abschnitt 3.2. führt TFIDF Vektoren als Grundlage für die Nutzung in Textklassifikation und E-Mail-Beantwortung ein. Abschließend beschreibt Abschnitt 3.3. die drei Anpassungen der Nächste-Nachbarn-Verfahren (NN-Verfahren), wie sie mit dieser Arbeit programmiert und in ihrer Leistungsfähigkeit untersucht wurden.

#### 3.1. Allgemeine Funktionsweise

Während der Trainingsphase speichert das NN-Verfahren die einzelnen Trainingsbeispiele  $t_i$  ab, es finden noch keine Berechnungen statt. Erst bei der Klassifikation eines neuen Beispiels  $t_q$  wird die eigentliche Rechenarbeit ausgeführt. Das Verfahren geht davon aus, dass alle Beispiele in einem  $n$ -dimensionalen Vektorraum  $V^N$  als Punkte dargestellt werden können. Der nächste Nachbar  $t_{NN}$  ist dann durch die größte Ähnlichkeit  $\operatorname{argmax}_i(\operatorname{sim}(t_q, t_i))$  definiert. Zur Bemessung der Ähnlichkeit kann z. B. der kleinste Euklidische Abstand zwischen zwei Beispielen verwendet werden. Wenn  $\{a_1(x), \dots, a_n(x)\}$  den Vektor eines Beispiels darstellt, ist der Euklidische Abstand zwischen zwei Beispielen als

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^n (a_r(x_i) - a_r(x_j))^2} \quad (3.1.1)$$

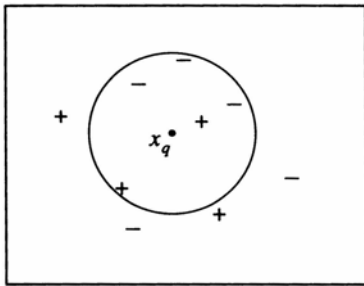
definiert.

Ist die Zielfunktion  $f(x)$  diskret, so wird aus den Klassen  $v \in V$  der  $k$  nächsten Nachbarn eine Mehrheitsentscheidung getroffen:

$$\hat{f}(x_q) \leftarrow \operatorname{argmax}_{v \in V} \left( \sum_{i=1}^k \delta(v, f(x_i)) \right), \quad (3.1.2)$$

$\delta(a, b) = 1$ , falls  $a = b$  und  $\delta(a, b) = 0$  sonst.

Das neu zu klassifizierende Element erhält das häufigste Klassenlabel. Dadurch ist es möglich, für unterschiedliche Werte von  $k$  auch unterschiedliche Klassifizierungsergebnisse für ein neues Element zu erhalten. Diese Situation ist in Abbildung 1 und der Verwendung von  $k=1$  sowie  $k=5$  verdeutlicht. Im ersten Fall klassifiziert das NN-Verfahren die Anfrage  $x_q$  als positiv und im zweiten Fall als negativ.



**Abbildung 1: Beispiel kNN-Methode [Mitchel, 1997]**

Kommen zur Trainingsmenge weitere Beispiele hinzu, so kann sich die Entscheidung für ein schon vorher klassifiziertes Beispiel wieder verändern, falls in der Nachbarschaft die Mehrheitsentscheidung durch die neuen Beispiele beeinflusst wird. Unter der Annahme, dass keine weiteren Trainingsbeispiele hinzukommen, lassen sich für einen bestimmten Wert von  $k$  Entscheidungstrennlinien zeichnen [iLink01]. Dadurch sind Regionen der Klassenzuordnung grafisch darstellbar. Für  $k=1$  ist jedes Trainingsbeispiel von einem konvexen Polyeder umrahmt. Liegen neu zu klassifizierende Beispiele innerhalb solch eines Polyeders, dann werden sie mit der Klasse des schon eingeschlossenen Beispiels markiert. Das Voronoi Diagramm für  $k=1$  ist in Abbildung 2 dargestellt.

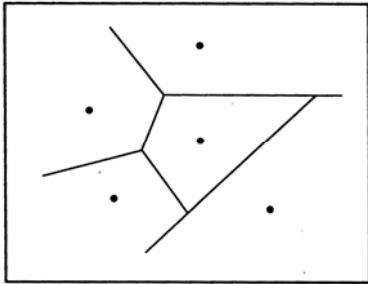


Abbildung 2: Voronoi Diagramm für  $k=1$  [Mitchel, 1997]

Vorab lässt sich keine Aussage darüber treffen, für welchen Wert  $k$  die besten Ergebnisse erzielbar sind. Für den jeweiligen Anwendungsfall müssen die verschiedenen Werte ausprobiert werden, um das beste Ergebnis herauszufinden. Allgemein kann festgestellt werden [iLink01], dass für kleine Werte von  $k$  eine hohe Sensibilität gegenüber Ausreißern besteht. Bei sehr hohen Werten von  $k$  zieht das Verfahren viele Elemente aus anderen Klassen hinzu.

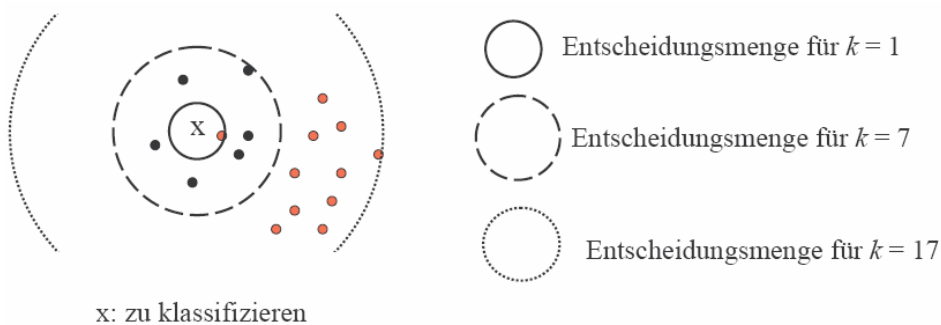


Abbildung 3: Unterschiedliche Ergebnisse der kNN-Methode bei Variation von  $k$  [iLink01]

### 3.2. Das Vektorraummodell

Um mit den NN-Verfahren Texte zu klassifizieren, müssen zuerst die Voraussetzungen dafür geschaffen werden. Wie im Abschnitt 3.1. aufgeführt, benötigen wir die Trainingsbeispiele als Punkte in einem  $n$ -dimensionalen Vektorraum  $V^N$ .

Die Grundlage hierfür ist das Vektorraummodell, wie in [Manning/Schütze, 1999] beschrieben. Hierbei sind Dokumente und Anfragen in einem hochdimensionalen Vektorraum abgebildet, wobei jede Dimension ein Wort aus der Textsammlung darstellt.

Für jedes Wort  $w_i$  im Dokument  $d_j$  bezeichnet die Termfrequenz  $tf_{i,j}$  die Anzahl der Wiederholungen des Wortes  $w_i$  im Dokument  $d_j$ . Die nächsten Nachbarn einer E-Mail liegen im Vektorraummodell sehr nahe am Anfragevektor. Bis hierher würden große Texte längere Vektoren erzeugen als inhaltlich identische, kürzere Texte. Somit könnte zwischen zwei solchen Texten nur schwerlich eine Ähnlichkeit festgestellt werden.

Anstatt die Größe des Vektors zu betrachten, ist es häufig günstiger, die Winkel zwischen den Vektoren als Entscheidungskriterium zu werten.

Der Kosinus zwischen zwei Vektoren ist definiert als

$$\cos(x_1, x_2) = \frac{\sum_{i=1}^n x_{i1} x_{i2}}{\sqrt{\sum_{i=1}^n x_{i1}^2} \sqrt{\sum_{i=1}^n x_{i2}^2}}. \quad (3.2.1)$$

Werden die Vektoren nun noch vorher normalisiert, d.h.  $\sqrt{\sum_{i=1}^n x_i^2} = 1$  gilt für jeden Vektor,

dann berechnet sich der Kosinus als

$$\cos(x_1, x_2) = \frac{x_1 * x_2}{\|x_1\| * \|x_2\|}. \quad (3.2.2)$$

Das jeweils nächste Element hat nun den kleinsten Winkel zur Anfrage. Für die Ähnlichkeit zwischen zwei Elementen gilt die Beziehung  $sim(x_1, x_2) = \cos(x_1, x_2)$ .

Nun stellt sich noch die Frage, wie die einzelnen Wörter und Elemente des Vektors gewichtet werden. Die Definition der Termfrequenz zählt bis jetzt das Auftreten eines Wortes im Text. Das würde bedeuten, ein dreimaliges Vorkommen gibt diesem Wort auch

eine dreimalig größere Relevanz für den speziellen Texttyp. Um diesen Effekt abzuschwächen, wird die Termfrequenz als  $tf_{i,j}=1+\log(\# w_{i,j})$  berechnet, wodurch sich die Bedeutung nur gering erhöht.

Die zweite wichtige Kennzahl ist die Dokumentenfrequenz  $df_i$ . Sie misst die Anzahl der Texte mit dem Wort  $w_i$ . Baut man diese zur inversen Dokumentenfrequenz  $idf_i = \log \frac{N}{df_i} = \log N - \log df_i$  um, so wird Wörtern, die in nur einem Dokument vorkommen, das maximale Gewicht  $idf_i = \log N - \log 1 = \log N$  zugewiesen. Wörter, die in jedem Text auftauchen, erhalten die Gewichtung  $idf_i = \log N - \log N = 0$ .

Für das TFIDF-Vektorraummodell werden nun die Termfrequenz und die inverse Dokumentenfrequenz kombiniert, um die einzelnen Wörter im hochdimensionalen Vektorraum zu bestimmen. Es entsteht die Formel für die TFIDF-Vektorraumrepräsentation

$$tfidf_{i,j} = (1 + \log(tf_{i,j})) \log \frac{N}{df_i} \quad \text{für } tf_{i,j} \geq 1 \quad (3.2.3)$$

und  $tfidf_{i,j}=0$ , falls  $tf_{i,j}=0$ . Nachdem jedes Element des Vektors auf diesem Weg ermittelt wurde, müssen die Vektoren noch normalisiert werden, um die einfache Berechnung des Kosinus zu gewährleisten.

### 3.3. Anpassungen an die E-Mail-Beantwortung

Die Datensätze bestehen aus E-Mail-Anfragen und den zugehörigen Antworten. In diesem Abschnitt stellen wir die drei durch uns entwickelten Ansätze zur NN-Klassifikation vor, die dann im 6. Abschnitt weiter untersucht werden. Wir wenden das NN-Verfahren sowohl mit den Fragen als auch mit den Fragen und Antworten an. Die E-Mails liegen schon in der Vektorraumrepräsentation vor.

### 3.3.1. NN-Klassifikation mit den Fragen

Abhängig von der Wahl des Parameters  $k$  werden zur Anfrage  $x_q$  aus den Trainingsdaten die nächsten Nachbarn  $\{x_1, \dots, x_k\}$  ermittelt. Das Element  $x_q$  entstammt den Testdaten. Nun kann aus den nächsten Nachbarn der Mittelwertsvektor mit der Formel

$$x_M = \frac{x_1 + \dots + x_k}{k} \quad (3.3.1.1)$$

berechnet werden, um damit eine Mehrheitsentscheidung abzubilden. Durch Iteration über alle Trainingsdaten erhalten wir den nächsten Nachbarn  $x_{NNM}$  mit der größten Ähnlichkeit zum Mittelwertsvektor:

$$x_{NNM} = \{x_i \mid \arg \max_{(x_i, y_i) \in M} \text{sim}(x_M, x_i)\}. \quad (3.3.1.2)$$

Für  $x_{NNM}$  muss  $x_{NNM} \in \{x_1, \dots, x_k\}$  gelten, da sich der Mittelwertsvektor innerhalb der nächsten Nachbarn befindet. Die Trainingsdaten liegen schon in klassifizierter Form vor. Daher ist die Klasse von  $x_{NNM}$  auslesbar und mit der Klasse von  $x_q$  zu vergleichen. Nur wenn beide Klassen übereinstimmen, kann für die Akzeptanzfunktion  $\text{akzept}(x, f(x)) = 1$  gelten. Für die Erstellung der PR-Kurven wird in jedem Durchlauf der Grenzwert  $\Theta$  schrittweise erhöht. Die Akzeptanzfunktion kann nur  $\text{akzept}(x, f(x)) = 1$  zurückgeben, wenn die Ähnlichkeit  $\text{sim}(x_{NNM}, x_q)$  zwischen  $x_{NNM}$  und  $x_q$  über dem Grenzwert  $\Theta$  des aktuellen Durchlaufes liegt. Es gilt somit für die Antwort einer neuen Anfrage  $x_q$ :

$$f_\Theta(x_q) = \{y_i \mid \arg \max_{(x_i, y_i) \in M} \text{sim}(x_M, x_i)\}, \quad (3.3.1.3)$$

falls  $\text{sim}(x_{NNM}, x_q) > \Theta$  ist. Andernfalls wird keine Antwort zurückgegeben.

### 3.3.2. NN-Klassifikation mit den Fragen und Antworten

Identisch zur NN-Klassifikation mit den Fragen ermitteln wir zur Anfrage  $x_q$  aus den Trainingsdaten die nächsten Nachbarn  $\{x_1, \dots, x_k\}$ , wobei der Parameter  $k$  die Anzahl der Nachbarn festlegt. Das Element  $x_q$  ist wieder den Testdaten entnommen. Im Unterschied zur Klassifikation mit den Fragen rechnen wir jetzt mit den Antworten  $\{y_1, \dots, y_k\}$  weiter. Es wird der Mittelwertsvektor unter den Antworten mit der Formel

$$y_M = \frac{y_1 + \dots + y_k}{k} \quad (3.3.2.1)$$

berechnet. Die Iteration über alle  $\{y_1, \dots, y_k\}$  sucht den nächsten Nachbarn  $y_{NNM}$  zum Mittelwertsvektor  $y_M$  und bildet damit die Mehrheitsentscheidung ab:

$$y_{NNM} = \{y_i \mid \arg \max_{(x_i, y_i) \in M} \text{sim}(y_M, y_i)\}. \quad (3.3.2.2)$$

Die Trainingsdaten liegen schon in klassifizierter Form vor. Demnach ist die Klasse von  $y_{NNM}$  bekannt und wird mit der Klasse von  $y_q$  verglichen. Nur wenn  $y_{NNM}$  und  $y_q$  der gleichen Klasse  $S_i$  entstammen, kann die Akzeptanzfunktion  $\text{akzept}(x, f(x)) = 1$  liefern.

Der für die Erstellung der PR-Kurven notwendige Grenzwert  $\Theta$  wird auf die Ähnlichkeit zwischen  $x_{NNM}$  und  $x_q$  angewendet. Damit die Akzeptanzfunktion  $\text{akzept}(x, f(x)) = 1$  liefert, muss die Ähnlichkeit zwischen  $x_{NNM}$  und  $x_q$  größer als der Grenzwert  $\Theta$  des aktuellen Durchlaufs sein. Es gilt somit für die Antwort einer neuen Anfrage  $x_q$ :

$$f_\Theta(x_q) = \{y_i \mid \arg \max_{(x_i, y_i) \in M} \text{sim}(y_M, y_i)\}, \quad (3.3.2.3)$$

falls  $\text{sim}(x_{NNM}, x_q) > \Theta$  ist. Andernfalls wird keine Antwort zurückgegeben.

### 3.3.3. NN-Klassifikation mit den Fragen und Antworten (Ähnlichkeitsmittelung der Fragen)

Die NN-Klassifikation mit den Fragen und Antworten, basierend auf der Ähnlichkeitsmittelung, arbeitet bis auf die Bildung des Grenzwertes identisch zum Abschnitt 3.3.2. Die Ähnlichkeit wird nach der Formel

$$sim = \frac{sim(x_q, x_1) + \dots + sim(x_q, x_k)}{k} \quad (3.3.3.1)$$

bestimmt und auf den PR-Kurven-Grenzwert  $\Theta$  des aktuellen Durchlaufes angewendet. Es gilt somit für die Antwort einer neuen Anfrage  $x_q$ :

$$f_{\Theta}(x_q) = \{y_i \mid \arg \max_{(x_i, y_i) \in M} sim(y_M, y_i)\}, \quad (3.3.3.2)$$

falls  $\frac{sim(x_q, x_1) + \dots + sim(x_q, x_k)}{k} > \Theta$  ist. Andernfalls wird keine Antwort zurückgegeben.

## 4. Referenzmethoden zur automatischen E-Mail-Beantwortung

Mit dieser Arbeit wurden die im Abschnitt 3.3. beschriebenen NN-Klassifikatoren für die automatische E-Mail-Beantwortung programmiert. Ihre Leistungsfähigkeit soll mit ähnlichen Methoden verglichen werden. Die überwachte Klassifikation (SV-Klassifikation) und das Lernen aus Frage-Antwort-Paaren (LP-Methode) sind als Referenzmethoden der Arbeit von [Bickel/Scheffer, 2004] entnommen und dort ausführlich beschrieben. Daher sollen beide im 4. Abschnitt nur kurz erläutert werden.

### 4.1. Überwachte Klassifikation

Die Arbeit [Bickel/Scheffer, 2004] benötigt für die automatisierte E-Mail-Beantwortung eine obere Leistungsgrenze und nutzt dazu die überwachte Klassifikation (SV-

Klassifikation). Die Trainingsdaten werden mit zusätzlichen Informationen versehen und manuell in semantisch äquivalente Klassen eingeteilt. Somit ist das E-Mail-Beantwortungsproblem in ein Klassifikationsproblem transferiert, da für neue E-Mails die entsprechende Äquivalenzklasse gesucht wird.

Die Klassen  $S = \{S_1, \dots, S_N\}$  sind manuell erstellt, so dass jede Klasse  $S_i$  semantisch äquivalente E-Mail-Paare enthält und  $S_i = \{(x_{i1}, y_{i1}), \dots, (x_{im}, y_{im})\}$  gilt. Jeder Klasse  $S_i$  wird eine Standardantwort  $y_i^*$  zugewiesen, damit für jede E-Mail  $(x_j, y_j) \in S_i$  gilt:  $(x_j, y_j) \in S_i \Rightarrow akzept(x_j, y_i^*) = 1$ . Die Klasse  $S_N$  enthält sonstige E-Mails, die keiner anderen Klasse zugeordnet werden können.

Es ist nun ein Klassifikator gesucht, der einer neuen E-Mail-Anfrage  $x$  eine Standardantwort  $f(x)=y_i^*$  zuweist.

Das Klassifikationsproblem wird in [Bickel/Scheffer, 2004] mit Support Vektor Maschinen (SVM) gelöst. Die  $N$  binären SVM arbeiten mit einem einer-gegen-alle-Ansatz, und zerlegen dazu die Daten in positive und negative Beispiele. Die positiven Beispiele für eine Klasse  $i$  sind dabei die eingehenden Nachrichten  $Pos_i = \{x_{i1}, \dots, x_{im}\}$  mit  $S_i = \{(x_{i1}, y_{i1}), \dots, (x_{im}, y_{im})\}$ . Die negativen Beispiele einer Klasse  $i$  sind alle  $Pos_j$  mit  $i \neq j$ .

Die weitere Anwendung der Support Vektor Maschinen auf die Äquivalenzklassen ist detailliert in [Bickel/Scheffer, 2004] beschrieben.

## 4.2. Lernen aus Frage-Antwort-Paaren

Bevor die SV-Methode eine neue E-Mail automatisch beantworten kann, sind die Trainingsbeispiele manuell zu klassifizieren [Bickel/Scheffer, 2004]. Dieser Schritt ist aufgrund des Zeitbedarfes ungeeignet für die Praxis. Um die manuelle Klassifizierung zu umgehen, sollen mit dem Lernen aus Frage-Antwort-Paaren (LP-Methode) die Klassen automatisch generiert werden.

Der Grundgedanke besteht darin, die manuelle Klassifikation durch einen automatisierten Clusterschritt zu ersetzen.

Der Clusteralgorithmus arbeitet auf den Antworten  $y_i$  der Trainingspaare  $(x_i, y_i)$  und erstellt die Klassen  $C = \{C_1, \dots, C_R\}$ . Ein Cluster  $C_i = \{(x_{i1}, y_{i1}), \dots, (x_{im}, y_{im})\}$  enthält E-Mail-Paare, deren Antworten  $y_i$  im gleichen Cluster  $C_i$  liegen. Der Cluster  $C_R$  enthält alle Elemente, die keinen hinreichend großen Cluster bilden konnten. Für jeden Cluster wird eine Standardantwort  $y_i^*$  mittels der Formel

$$y_i^* = \arg \max \left( \text{sim} \left( y, \frac{1}{|C_i|} \sum_{(x_j, y_j) \in C_i} y_j \right) \right) \quad (4.2)$$

definiert, wobei  $\arg \max$  über alle  $y : (x, y) \in C_i$  iteriert. Damit liegt die Standardantwort für einen Cluster in der Nähe des Centroids.

Die Clusterung wird mit dem EM-Algorithmus ausgeführt. Anschließend werden in der Arbeit von [Bickel/Scheffer, 2004] auf den Fragen R binäre, lineare Support Vektor Maschinen (SVM) trainiert. Die SVM arbeiten hier identisch wie im Abschnitt 4.1. beschrieben. Die darüber hinausgehenden detaillierten Ausführungen zur Clusterung und den SVM sind in der Arbeit [Bickel/Scheffer, 2004] enthalten.

## 5. Datensätze zur Evaluierung

Die Leistungsfähigkeit der NN-Klassifikatoren soll an zwei Testdatensätzen überprüft werden. Es stehen die Daten eines großen Online-Versandhändlers sowie einer Krankenhaus-EDV-Hotline zur Verfügung. Zuerst werden die Testdatensätze beschrieben und danach auf ihre Tauglichkeit zur automatischen E-Mail-Beantwortung überprüft. Da wir die NN-Klassifikatoren hinsichtlich eines Praxiseinsatzes untersuchen, bedeutet Tauglichkeit für uns, eine akzeptable obere Leistungsgrenze festzustellen. Wie im Abschnitt 4.1. beschrieben und in [Bickel, Scheffer, 2004] verwendet, nutzen wir die SV-Klassifikation zur Identifikation der oberen Leistungsgrenze.

## 5.1. Online-Versandhändler

Die Daten wurden von einem großen Online-Versandhändler bereitgestellt. Es sind 805 Frage-Antwort-Paare enthalten, die von einem Mitarbeiter der Hotline in einem Monat geschrieben wurden. Die meisten Anfragen behandeln Themen wie verspätete, unvollständige oder defekte Lieferungen. Um die Leistung der Verfahren messen zu können, haben wir die E-Mail-Paare manuell in 19 semantisch äquivalente Klassen eingeteilt. Von den 805 Paaren konnten 203 Paare in keine Klasse eingeordnet werden oder ihrerseits keine ausreichend große Klasse bilden. Diese 203 Paare bezeichnen wir als sonstige E-Mails.

Entsprechend unseren vorab durchgeführten Auswertungen und den Ergebnissen in [Bickel/Scheffer, 2004] erzielen die NN-Klassifikatoren durch die Verwendung der sonstigen E-Mails schlechtere Leistungswerte. Wir verwenden daher in den Auswertungen nur die 602 E-Mail-Paare, die wir manuell einer Klasse zuordnen konnten. Die 203 sonstigen E-Mails sind in den folgenden Auswertungen nicht betrachtet. Da mit der Arbeit [Bickel/Scheffer, 2004] ebenso Auswertungen ohne die sonstigen E-Mails ausgeführt wurden, können wir die Vergleichbarkeit der Methoden gewährleisten.

### 5.1.1. Verifikation der Tauglichkeit

Die Tauglichkeit wird hier über die Ergebnisse in [Bickel/Scheffer, 2004] bestätigt. Wie im 2. Abschnitt aufgeführt, verwenden wir die PR-Kurven, um die Leistungsfähigkeit der NN-Klassifikatoren zu bewerten. Für einen Praxiseinsatz ist uns wichtig, dass der Datensatz eine akzeptable obere Leistungsgrenze aufweist. Hierfür verwenden wir die SV-Klassifikation, wie in Abschnitt 4.1. beschrieben und in [Bickel, Scheffer, 2004] verwendet.

Auf den Daten des Onlinehändlers konnte die SV-Klassifikation mit zufrieden stellenden Ergebnissen angewendet werden und ist in Abbildung 4 dargestellt. So werden z. B. bei einer Beantwortungsrate von 25% immerhin 50% der Antworten akzeptabel beantwortet.

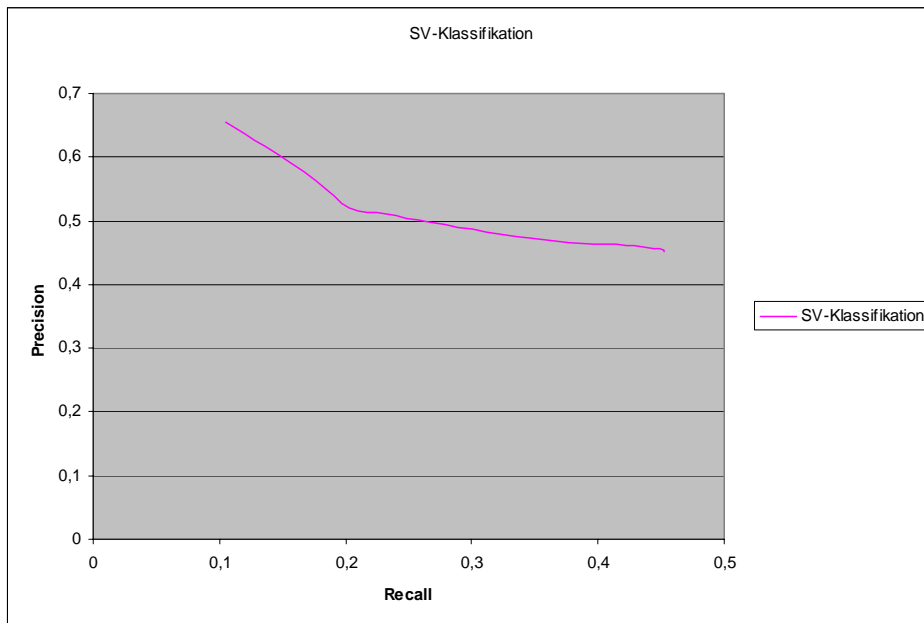


Abbildung 4: SV-Klassifikation der Daten des Online-Versandhändlers

## 5.2. EDV-Hotline im Krankenhaus

Die Daten stammen aus der EDV-Hotline eines großen Krankenhauses. Es liegen 1600 gesendete E-Mails vor, die von einer Mitarbeiterin der EDV-Hotline in einem Jahr geschrieben wurden.

Nach der manuellen Sichtung war erkennbar, dass ca. 50% aller gesendeten E-Mails an Kollegen verschickt wurden und somit für eine automatisch generierte Antwort nicht in Frage kamen. Entweder war die vorherige Anfrage zu projektbezogen oder es gab keine schriftliche Anfrage, da vorab mündlich kommuniziert wurde. Unter Zuhilfenahme einer Mitarbeiterliste konnten wir an Kollegen verschickte E-Mails aussortieren.

Für die verbleibenden 800 E-Mails sollte sichergestellt werden, dass Frage-Antwort-Paare vorliegen. Da der verwendete E-Mail-Client beim Antworten den String „Ursprüngliche Nachricht“ einfügt, konnten die 378 Frage-Antwort-Paare leicht erkannt werden.

Somit sind von den 1600 bereitgestellten E-Mails nur 24% potentiell automatisch beantwortbar.

Um die im 2. Abschnitt beschriebenen Grundlagen zu schaffen, mussten nun die 378 E-Mails manuell in Klassen eingeteilt werden. Dabei konnten nur 47% in Klassen mit mehr als 8 Elementen einsortiert werden. 200 E-Mails waren keiner Klasse zuweisbar und sind im Folgenden als sonstige E-Mails bezeichnet. Dadurch sind nun von den 1600 versendeten E-Mails nur noch 13% potentiell automatisch beantwortbar.

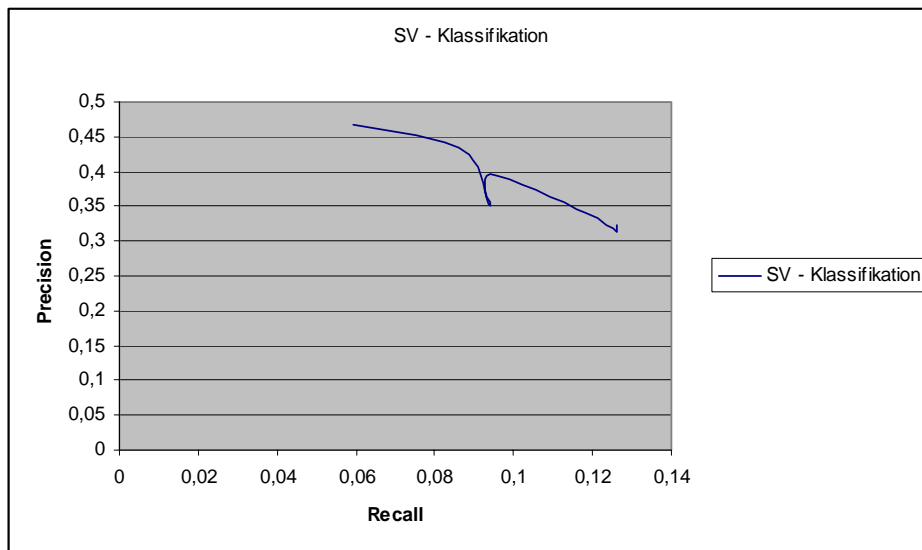
1.	Erweiterung der E-Mail-Postfachgröße
2.	Änderung von Berechtigungen auf einem Netzlaufwerk
3.	PC-Umstellung/ -Umzug
4.	Neueinrichtung eines Netzlaufwerkes
5.	Anfrage muss durch <u>IT-Leiter</u> beantwortet werden
6.	Einrichtung eines Druckers
7.	Erhöhung des Speicherplatzes auf dem Netzlaufwerk
8.	Neueinrichtung eines Nutzers
9.	Freischaltung von PC-Laufwerken
10.	Einrichtung eines E-Mail-Postfaches
11.	Wiederherstellung von Dateien
12.	Sonstige E-Mails

**Abbildung 5: Klasseneinteilung der EDV-Hotline-E-Mails**

### 5.2.1. Verifikation der Tauglichkeit

Wie auch im Abschnitt 5.1.1. soll die obere Leistungsgrenze für die PR-Kurven bestimmt werden, um damit die Praxistauglichkeit der Daten zu bewerten.

Die SV-Klassifikation wird auf die verbleibenden 13% der gesendeten E-Mails angewendet und liefert das in Abbildung 6 gezeigte Ergebnis.



**Abbildung 6: SV-Klassifikation der EDV-Hotline-Daten**

Diese 13% enthalten nur E-Mail-Paare, die wir manuell einer Klasse zuordnen konnten. Entsprechend unseren vorab durchgeführten Auswertungen und den Ergebnissen in [Bickel/Scheffer, 2004] ergibt sich unter Einbeziehung der sonstigen E-Mails eine noch schlechtere Leistungskurve.

So kann mit den Daten der EDV-Hotline nach Abbildung 6 z. B. bei einer Beantwortungsrate von 13% nur eine Genauigkeit von 31% erzielt werden.

Da nur noch 13% aller E-Mails potentiell automatisch beantwortbar sind und zusätzlich die Beantwortung dieser 13% eine sehr schlechte Leistung aufweist, wird der Datensatz für die automatische E-Mail-Beantwortung als untauglich eingestuft und nicht weiter verwendet.

## 6. Auswertung

In diesem Abschnitt sollen die Ergebnisse der verschiedenen NN-Klassifikationen vorgestellt und anschließend mit ähnlichen Methoden verglichen werden.

### 6.1. Anwendung der NN-Klassifikatoren

Wie im Abschnitt 3.3. beschrieben, wollen wir die drei Anpassungen der NN-Klassifikation auf die Testdaten anwenden. Der Datensatz des Krankenhauses ist aufgrund

der im Abschnitt 5.2.1. entwickelten Argumente nicht verwendbar. Alle folgenden Diagramme beziehen sich somit auf die Daten des Online-Versandhändlers. Für den Vergleich der NN-Verfahren benötigen wir keine sonstigen E-Mails.

### 6.1.1. NN-Klassifikation mit den Fragen

Die kNN-Klassifikation mit den Fragen und dem Parameter  $k=1$  wurde auch schon in [Bickel/Scheffer, 2004] angewendet. Hier wollen wir das Verhalten für steigende  $k$ -Werte untersuchen. In Abbildung 7 ist die zugehörige PR-Kurve abgebildet.

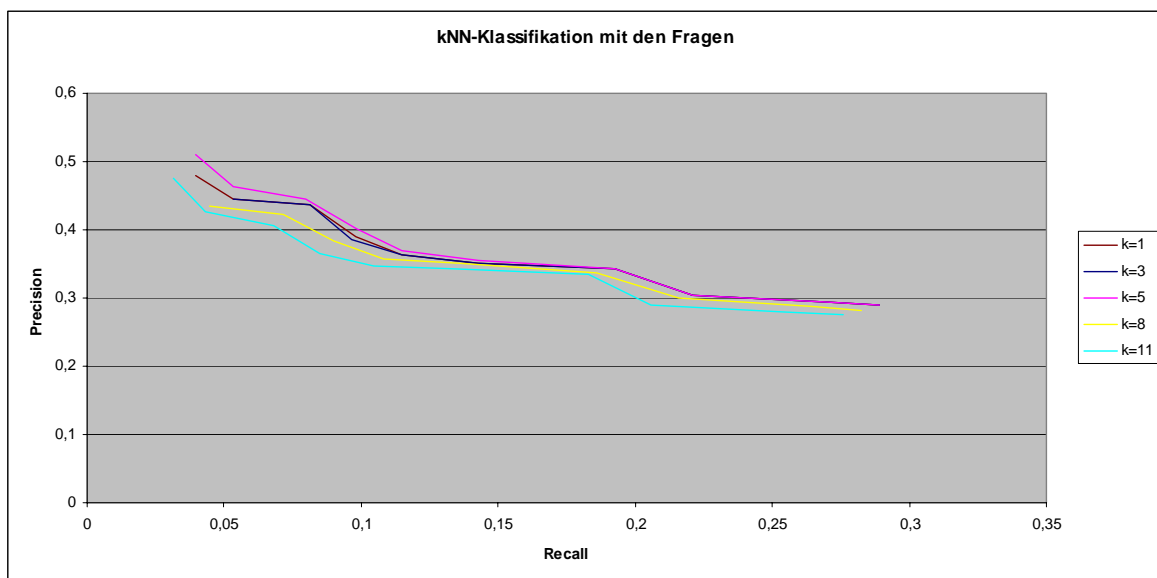


Abbildung 7: kNN-Klassifikation mit den Fragen

Mit der Erhöhung von  $k$  steigt die Leistung in kleinen Schritten an und findet bei  $k=5$  ihren maximalen Punkt. So kann bei einer Beantwortungsrate von 28% eine Genauigkeit von 30% erzielt werden, wodurch wir gegenüber dem 1NN-Verfahren einen Erfolg erreichen konnten.

Begründbar ist die Verbesserung durch das Hinzuziehen weiterer Informationen aus der Umwelt mit steigendem Parameter  $k$ .

Wie im Abschnitt 3.1. erläutert, ist das Verfahren bei der Verwendung von kleinen  $k$ -Werten anfällig gegen Störungen. Mit der Erhöhung von  $k$  wird auch die Stabilität

angehoben. Das beste Ergebnis erreichen wir mit dem Parameter  $k=5$ . Mit noch weiter steigenden  $k$ -Werten fällt sie wieder ab, da Objekte aus fremden Klassen betrachtet werden.

### 6.1.2. NN-Klassifikation mit den Fragen und Antworten

Wie im Abschnitt 3.3.2. aufgeführt, untersucht dieser NN-Klassifikationsansatz die nächsten Nachbarn auf der Antwortseite.

Für den Wert  $k=1$  muss sich das gleiche Ergebnis wie im Abschnitt 6.1.1. ergeben, da nur eine Nachbarfrage  $x_1$  zur Verfügung steht. Der Mittelwert  $y_{NNM}$  auf der Antwortenseite wird unter den verfügbaren Antworten  $\{y_1, \dots, y_k\}$  ausgewählt und muss daher auf die eine Antwort  $y_1$  fallen.

Abbildung 8 zeigt die PR-Kurve der NN-Klassifikation mit den Fragen und Antworten.

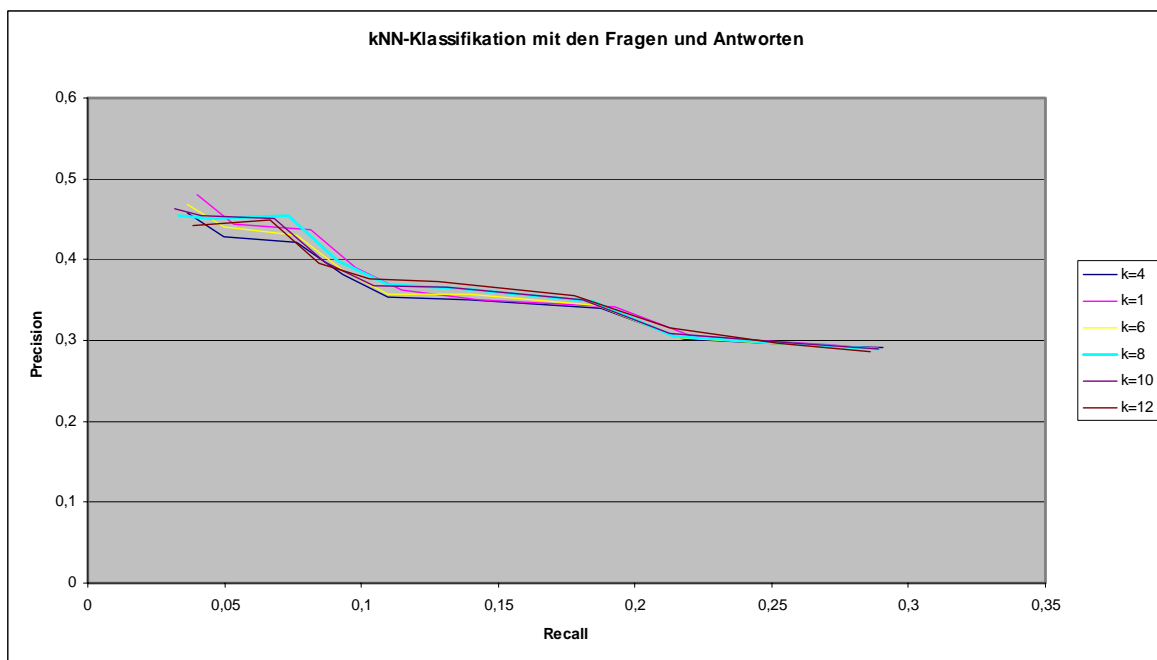


Abbildung 8: kNN-Klassifikation mit den Fragen und Antworten

Auffällig ist, dass für  $k=1$  ein vergleichsweise guter Leistungswert erreicht werden konnte. Das steht im Gegensatz zur beschriebenen Theorie im Abschnitt 3.1.. Begründbar ist dieses Verhalten durch das abgeänderte Grundprinzip der NN-Klassifikation mit den Antworten.

Die nächsten Nachbarn werden auf der Frageseite bestimmt, dagegen der Mittelpunkt  $y_M$  und das nächste Element  $y_{NNM}$  auf der Antwortseite.

Mit steigendem Wert von  $k$  fällt die Leistung ab, um dann ab  $k=6$  wieder zu steigen. Das globale Maximum ist bei  $k=8$  erreicht, wonach die Leistung wiederum leicht abfällt. Dieser Abfall ist konform zur oben beschriebenen Theorie.

### 6.1.3. NN-Klassifikation mit den Fragen und Antworten (Ähnlichkeitsmittelung der Fragen)

Mit diesem Ansatz wird die Ähnlichkeit der nächsten Fragenachbarn gemittelt. Bei der Anwendung des Grenzwertes  $\Theta$  auf diese Ähnlichkeit muss die Leistung gleich oder schlechter sein als im Abschnitt 6.1.2., da für alle  $\text{sim}(x_q, x_i)$  gilt:

$$\text{sim}(x_q, x_{NNM}) \geq \text{sim}(x_q, x_i), \text{ für } x_q \neq x_{NNM}. \quad (6.1.3.1)$$

Daraus folgt

$$\text{sim}(x_{NNM}, x_q) \geq \frac{\text{sim}(x_q, x_1) + \dots + \text{sim}(x_q, x_k)}{k}. \quad (6.1.3.2)$$

Die Formeln 3.3.3.1 und 3.3.3.2 geben nur dann eine Antwort zurück, wenn im ersten Fall  $\text{sim}(x_{NNM}, x_q) > \Theta$  und im zweiten Fall  $\frac{\text{sim}(x_q, x_1) + \dots + \text{sim}(x_q, x_k)}{k} \geq \Theta$  gilt. Es werden somit durch die Formel 3.3.3.2 akzeptable Antworten verworfen, die mit 3.3.3.1 noch verwendet worden wären. Daraus resultieren kleinere Precision- und Recall-Werte.

Daher muss die Ähnlichkeitsmittelung ein gleich gutes oder schlechteres Ergebnis erzielen als der Ansatz im Abschnitt 6.1.2. unter Benutzung der Ähnlichkeit zwischen  $x_q$  und  $x_{NNM}$ .

Abbildung 9 zeigt die zugehörigen PR-Kurven.

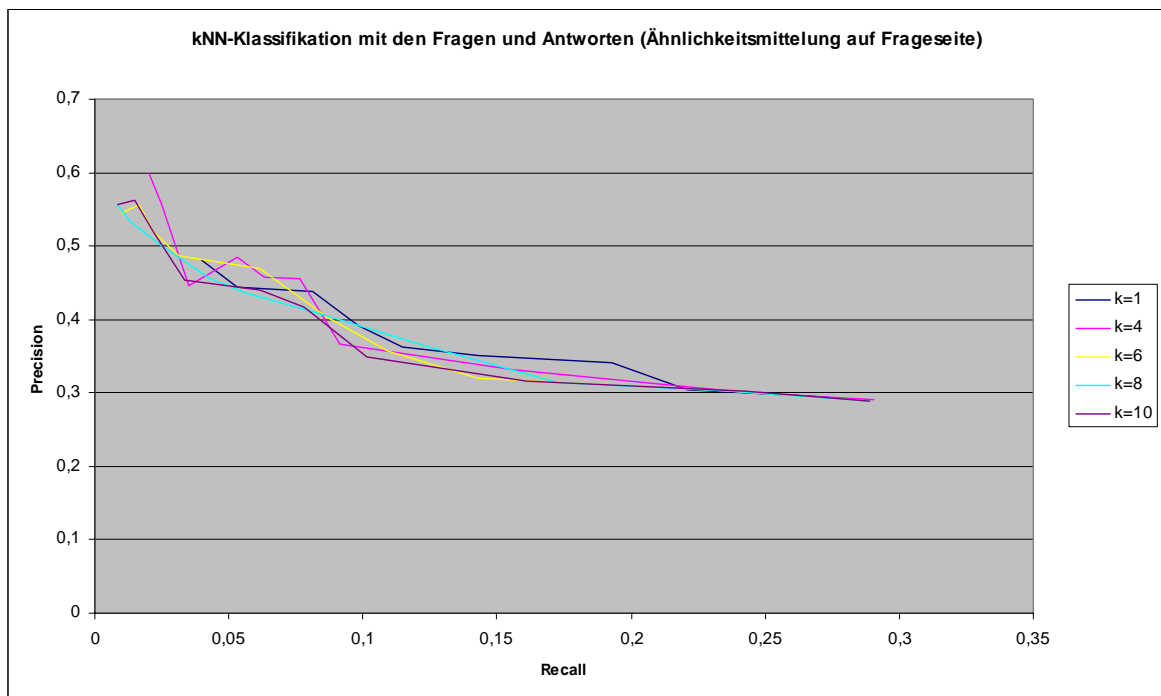


Abbildung 9: kNN-Klassifikation mit den Fragen und Antworten (Ähnlichkeitsmittelung Fragen)

Wie erwartet, wurde der beste Leistungswert mit dem Parameter  $k=1$  gewonnen. Für steigende Werte von  $k$  fällt die Leistung ab und liegt damit unter den Ergebnissen aus Abschnitt 6.1.2.

## 6.2. Vergleich ähnlicher Methoden

Die NN-Klassifikatoren wollen wir mit den im 4. Abschnitt eingeführten Methoden SV-Klassifikation und LP-Klassifikation vergleichen.

Abbildung 10 stellt den Vergleich der einzelnen E-Mail-Beantwortungsverfahren dar, wobei die sonstigen E-Mails für den Vergleich nicht betrachtet werden.

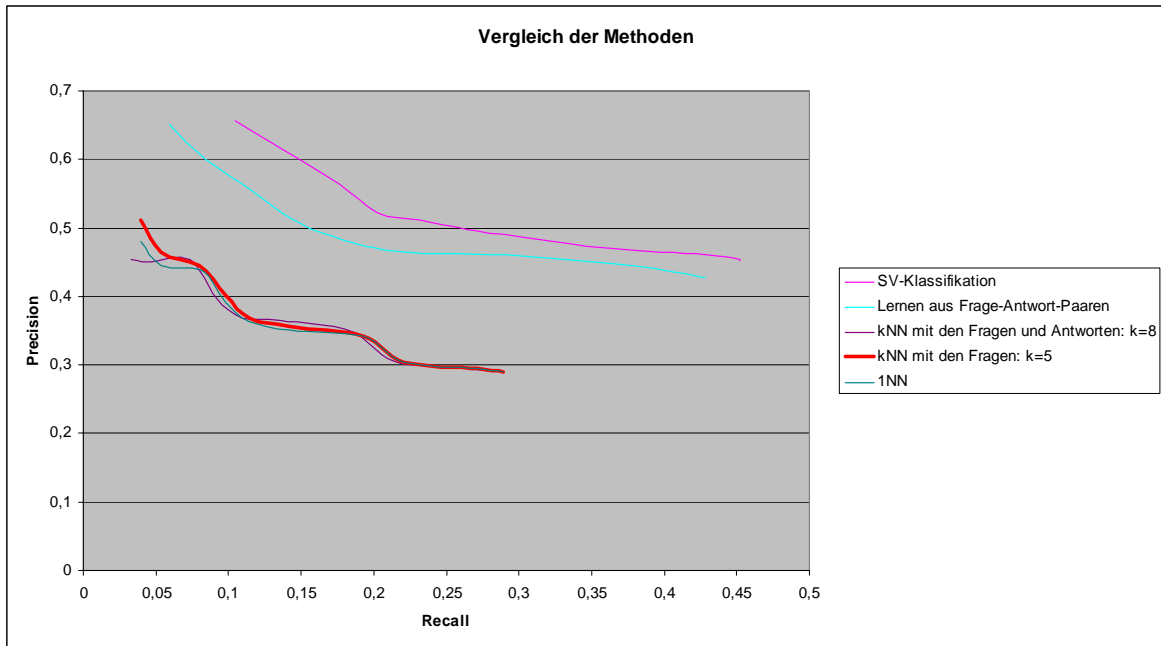


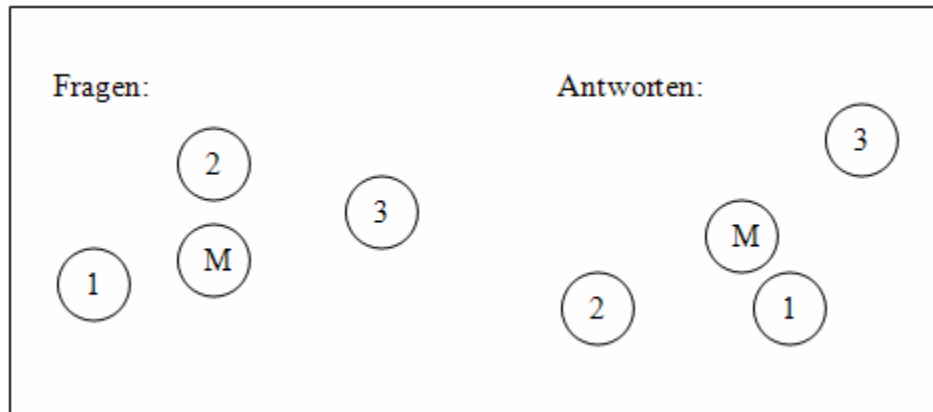
Abbildung 10: Vergleich verschiedener Methoden zur E-Mail-Beantwortung

Die besten Ergebnisse werden durch die SV-Klassifikation erzielt. Wie schon in [Bickel/Scheffer, 2004] erläutert, ist sie eine obere Grenze für die PR-Kurven, da durch die vorherige manuelle Klassifikation zusätzliche Informationen vorliegen.

Leistungsmäßig in der Mitte liegt das Lernen aus Frage-Antwort-Paaren. Die NN-Klassifikatoren sind im unteren Leistungsbereich angesiedelt, wie auch schon in [Bickel/Scheffer, 2004] angesprochen.

Von den drei mit dieser Arbeit programmierten NN-Klassifikatoren hat sich die NN-Klassifikation auf den Fragen als am besten herausgestellt. Innerhalb dieser Methode konnten wir beste Ergebnis mit dem Parameter  $k=5$  erzielen.

Bezogen auf das gesamte Leistungsspektrum gibt es zwischen den einzelnen NN-Klassifikatoren keine bedeutenden Leistungsunterschiede. Das doch Unterschiede in der Theorie existieren, verdeutlicht Abbildung 11.



**Abbildung 11: Unterschiedliche Mittelpunkte bei Anwendung der NN-Klassifikation mit den Fragen und Antworten**

Auf der linken Seite sind die Fragen und auf der rechten Seite die zugehörigen Antworten beispielhaft im Vektorraummodell abgebildet. Das Element M stellt jeweils den Mittelpunkt dar. Unter den Fragen liegt E-Mail Nr. 2 in der Nähe vom Mittelpunkt, wogegen unter den Antworten E-Mail Nr. 1 die größte Ähnlichkeit aufweist.

Damit ist zu erklären, warum die im Abschnitt 3.3. erläuterten Methoden unterschiedliche Leistungswerte aufweisen.

## 7. Schlussfolgerungen

Wir haben mit dieser Arbeit Methoden untersucht, um E-Mails automatisch beantworten zu können. Dazu wurden drei NN-Klassifikatoren programmiert und auf eine Fallstudie angewendet.

Zusätzlich konnten wir die NN-Klassifikatoren mit ähnlichen E-Mail-Beantwortungsverfahren aus [Bickel/Scheffer, 2004] vergleichen.

Unter den drei programmierten NN-Ansätzen hat sich die Suche in den Fragen als am leistungsfähigsten herausgestellt. Die besten Ergebnisse konnten wir bei Betrachtung der 5 nächsten Fragen erzielen.

Der Vergleich der NN-Klassifikation mit ähnlichen Methoden hat die Ergebnisse aus [Bickel/Scheffer, 2004] bestätigt. In der Suche nach automatischen E-Mail-Beantwortungsmethoden stellen die NN-Klassifikatoren eine untere Leistungsgrenze dar. Die NN-Verfahren ziehen, im Gegensatz zu ähnlichen Methoden, weniger Informationen aus der Umwelt hinzu. Dieser Fakt kann als Ursache für die geringere Leistungsfähigkeit gesehen werden.

Für den praktischen Einsatz zur automatischen Beantwortung von E-Mails sollte das in [Bickel/Scheffer, 2004] entwickelte Lernen aus Frage-Antwort-Paaren benutzt werden. Erstens erreicht es im Vergleich zu den NN-Klassifikatoren bessere Leistungswerte. Zweitens entledigt es den Nutzer von der vorherigen manuellen Klassifizierung der Testdaten, wie es mit der SV-Klassifikation notwendig ist.

Aus den beiden Testdatensätzen lässt sich weiterhin erkennen, dass die Anwendungsdomäne von großer Bedeutung ist. Die automatische Beantwortung von E-Mails erfordert ähnliche Anfragen, wie sie in Servicecentern üblich sind. Diese Arbeit hat auch gezeigt, dass die EDV-Hotline des Krankenhauses noch nicht für den Einsatz automatisierter E-Mail-Beantwortungsverfahren geeignet ist. Daraus lässt sich ableiten, dass vor dem praktischen Einsatz Untersuchungen durchzuführen sind, um Aussagen über die Eignung einer Domäne treffen zu können. Ein einfacher Tauglichkeitstest wäre hier von großem Vorteil.

## Literatur- und Quellenverzeichnis

- [Bickel/Scheffer, 2004] Steffen Bickel, Tobias Scheffer: Learning from Message Pairs for Automatic Email Answering, in: *Proceedings of the European Conference on Machine Learning*, 2004.
- [Kockelkorn/Lüneburg/Scheffer, 2003] Michael Kockelkorn, Andreas Lüneburg, Tobias Scheffer: Learning to answer emails, in: *Proceedings of the International Symposium on Intelligent Data Analysis*, 2003
- [Mitchel, 1997] Tom M. Mitchel: Machine Learning, *McGraw-Hill*, 1997
- [Manning/Schütze, 1999] Christopher D. Manning, Hinrich Schütze: Foundations of statistical natural language processing, *MIT Press*, 1999
- [iLink01] Nächste-Nachbarn-Klassifikatoren (dbs.informatik.uni-muenchen.de),  
verfügbar: <http://www.dbs.informatik.uni-muenchen.de/Lehre/KDD/WS0304/Skript/kdd-3-klassifikation2.pdf> (zugegriffen: 1. 12. 2004)
- [Cohen, 1996] W. Cohen: Learning rules that classify email, in: *Proceedings of the IEEE Spring Symposium on Machine Learning for Information Access*, 1996
- [Sahamie et. al, 1998] M. Sahami, S. Dumais, D. Heckermann, E. Horvitz: A Bayesian Approach to filtering junk email, in: *Proceedings of the AAAI Workshop on Learning for Text Categorization*, 1998
- [Ducker et. al, 1999] H. Ducker, D. Wu, V. Vapnik: *Support vector machines for spam categorization*, in: *IEEE Transactions on Neural Networks*, 1999
- [Boone, 1998] T. Boone: Concept features in Re:Agent, an intelligent email agent, in: *Proceedings of the 2<sup>nd</sup> Annual Conference on Autonomous Agents*, 1998