Die Erkennung der deutschsprachigen Spam-Emails: Naïve-Bayes vs. ähnlichkeitsbasiertes Lernen

Sophia Katrenko
Universität Tübingen
Deutschland
katrenko@sfs.uni-tuebingen.de

Abstract

Das Filtering der Spam-Emails ist derzeit ein Problem, das noch zu lösen ist. Im Gegensatz zu früheren Arbeiten, die auf englischsprachigen Mails basieren, fokussiert dieser Artikel die deutschen Daten.Die Aufgabe der Spam-Erkennung wird als eine Klassifikationsaufgabe dargestellt, für deren Lösung die Lernverfahren verwendet werden. Es wird gezeigt, dass diese Methoden mit dem Englischen vergleichbare Ergebnisse liefern.

1 Einleitung

Die Erkennung von SpamNachrichten wurde in den letzten Jahren mehrmals betrachtet. Gewöhnlich ist Spam eine Bezeichnung für unerwünschte, oft kommerzielle Emails. Laut der Umfrage von www.marketagent.com bekommt ein deutschsprachiger Benutzer etwa 2 unerwünschte Nachrichten pro Tag, jeder vierte Internet-Nutzer bekommt mehr als 20 Massen-Emails wöchentlich. Die am meisten verbreiteten Themengebiete der Mails sind Erotik, Online-Shopping-Angebote und Kreditvermittlungen.

Die bisherigen Forschungen in dem Bereich des Spam-Filterings befassen sich fast ausschließlich mit den englischsprachigen Mails. Dabei werden Naïve-Bayes-Verfahren, Support Vector Machines (SVMs) und TF-IDF Algorithmus eingesetzt. Die erhaltenen Ergebnisse sind aber kaum vergleichbar, da die Forscher verschiedene Datenmengen verwendet haben. Im Grunde genommen können alle Forschungen in zwei Gruppen aufgeteilt werden: in die Klassifikation und das Clustering der Emails [Manco *et al.*, 2002]. Fast alle Experimente beruhen auf der Klassifikation, da die Clusteranalyse ein unüberwachtes Verfahren darstellt, das die automatische Kategorienbildung voraussieht. Beim Einsatz des Klassifikationsverfahrens werden alle Instanzen hingegen vordefinierten Klassen zugeordnet. Man sollte aber auch erwähnen, dass man die Klassifikation auf zwei Arten durchführen kann: die allgemeine Klassifikation der Emails und das Filtering der Spam-Nachrichten. Die besten Ergebnisse, die man für englischsprachige Emails bekommen hat, variieren von 95% der Genauigkeit während der allgemeinen Klassifikation bis 99% Genauigkeit für das SpamFiltering.

Die Ansätze der Spam-Klassifikation lassen sich auf folgende Weise unterscheiden:

- Qualität der Daten (Themenbereiche, Datenknappheit (engl: data sparseness) sowie auch die Zahl der vorhandenen Daten)
- Art der Information, die in Betracht gezogen wird (strukturelle, textuelle, usw.)

Als Beispiel könnte man die Filter nennen, welche häufig auf der Basis der Regel 'if-then' funktionieren, dabei nur die Information aus dem Kopfteil der Email berücksichtigt wird. Wie einige Forschungen gezeigt haben [Androutsopoulos *et al.*, 2000], sind diese Filter nicht sehr leistungsfähig. Somit ist festzustellen, dass die von den Filtern benutzte Information nicht ausreichend ist.

Es ist zu erwarten, dass die Zahl der Spam-Nachrichten in den anderen Sprachen erhöht wird, was einen Grund darstellt, mehr mit den deutschen Daten zu experimentieren.

In diesem Beitrag soll gezeigt werden, dass maschinelles Lernverfahren erfolgreich für die Erkennung der deutschsprachigen Spam-Emails verwendet werden kann. Diese Untersuchung unterscheidet sich von den anderen dadurch, dass sie Mails aus verschiedenen thematischen Bereichen berücksichtigt. Die gewählten Daten sollen zu mehreren Themenbereichen gehören, was den Nachrichtenarten des durchschnittlichen Internet-Nutzers entspricht. Auch beschränkt sich der unten beschriebene Ansatz nur auf das Spam-Filtering. Einerseits richtet sich das Interesse des Benutzers auf die Erkennung und das Entwerfen der unerwünschten Emails, andererseits ordnet jeder Internet-Nutzer die Emails nach seinen eigenen Prioritäten. Aufgrund dieser Fakten wird keine weitere Klassifikation in den Spam-Subkategorien sowie in den Kategorien, die normalerweise den Foldern des Email-Clienten entsprechen, durchgeführt.[1]

Im ersten Abschnitt wird beschrieben, wie man das Spam-Filtering als ein Klassifikationsproblem darstellen kann. Der zweite Abschnitt befasst sich mit den Daten und Methoden. Der dritte Abschnitt erläutert die Experimente. Der Artikel schließt mit einer Diskussion und gibt einen Ausblick.

2 Spam-Erkennung als eine Klassifikationsaufgabe

Die Aufgabe besteht darin, die Emails in zwei Kategorien zu klassifizieren, in Spam und Nicht-Spam. Mathematisch lässt sich das wie folgt beschreiben [Marquez, 2000] :

Es sei $T=\{ (x_1 , f(x_1)),...,(x_i , f(x_i))\}$ eine Menge der zu klassifizierenden Instanzen (Muster), wobei $x_i = x_{i1} , x_{i2} ,..., x_{im}$ die Attribute (Features) der Instanz x_i darstellen. Dann ist es das Ziel, eine unbekannte, in dem Eingaberaum

[1] Es wäre sinnvoll, wenn man die von einer Person erhaltenen Meldungen betrachtet

I definierte Funktion *f* durch T zu approximieren. Der Ausgaberaum 1,...,K sei diskret und ungeordnet. Somit ist die Funktion f als ein Mapping em f : I ->{1,...,K} anzusehen.

Der Lernalgorithmus produziert einen Klassifikator (engl.: classifier), der eine Hypothese über die Funktion *f* bietet.

Für die hier angegebene Aufgabe sind die Emails als Instanzen zu bezeichnen, deren Attribute die Elemente des Inhalts (engl.: body) von den Meldungen sind. Außerdem wird das Subject der Email zu den Attributen hinzugefügt.

3 Daten und Methoden

Der Korpus enthält 125 Spam- und 425 Nicht-Spam-Emails. 50,4% der unerwünschten Nachrichten sind aus dem Bereich 'Erotik' und der Rest bezieht sich auf 'Online-Shopping Angebote' (Urlaubsmöglichkeiten, Software, Übersetzungen usw.). Die Themenbereiche der Nicht-Spam-Emails entsprechen den Themen der Spam-Meldungen und werden aus deutschen Newsgroups (de.etc.finanz, de.comp, usw.) gesammelt. Einige Meldungen wurden auch von Privatpersonen erhalten.

Wie schon erwähnt, werden zwei Verfahren, das naïve Bayes und die beispiel-basierte Methode, verwendet. Als Software sind Timbl und Rainbow [Daelemans *et al.*, 2002; McCallum, 1996] zu benutzen.

Das naïve-Bayes-Verfahren basiert auf der Formel von Bayes für bedingte Wahrscheinlichkeiten. Da es eine naïve Annahme über die Unabhängigkeit der Attribute gebraucht, wurden beim Einsatz dieses Verfahrens in den vielen früher durchgeführten Experimenten gute Ergebnisse erzielt.

Das Verfahren des nächsten Nachbarn repräsentiert die ähnlichkeitsbasierten (beispiel-basierten) Methoden. Hierbei handelt es sich darum, dass zwei Instanzen x_1 und x_2 die Nachbarn genannt werden, wenn die Instanz x_1 unter den k nächsten Nachbarn x_2 gehört und vice versa. Es soll betont werden, dass es verschiedene Massen gibt, mittels welche die Ähnlichkeit zu errechnen ist. Auch existieren mehrere Möglichkeiten, die Gewichtungen den Attributen zuzuweisen, darunter 'information gain', chi-squared usw.

4 Experimente

4.1 Datenverarbeitung

Die gewöhnlichen Vorverarbeitungsschritte bestehen aus Tokenization, Stemming (Stammformreduktion) und Entfernen der häufig vorkommenden Wörter.

Unter Tokenization versteht man das Extrahieren der Tokens (Wörter), wobei man die Interpunktion von den Wörtern trennt.

Beim Entfernen der Wörter wurden nur einige Einträge weggeworfen. Die englischen Stopp-Listen können beispielsweise auch Pronomen enthalten, die aber in diesem Ansatz aufgrund der Emailsberücksichtigung behalten wurden. Die Untersuchung der Emails zeigte beispielsweise, dass in den Spam-Emails eine höfliche Anrede sehr häufig vorkommt, während das bei den persönlichen Mails nicht der Fall ist. [2] Aus diesem Grund wurden nur Wörter wie 'www' und einige Präpositionen entfernt. Außerdem wurden alle Wörter gelöscht, die nur einen Buchstaben enthalten (dies kann erscheinen, wenn die Einträge zerlegt wurden).

Das Stemming ist in der Regel die Reduktion der Wörter auf ihre Wortstämme durch das Entfernen von Affixen. Es funktioniert gut im Englischen, aber führt zu Problemen, wenn man mit Deutsch arbeitet. Erstens charakterisiert sich Deutsch durch die grosse Anzahl der komplexen Wörter und zweitens können die Wortstämme oft geändert werden. Dies hat sich während der Forschung von [Hedstroem *et al.*, 2003] erwiesen, wo einen Stemmer gebraucht wurde, dessen Leistungsfähigkeit niedriger als erwartet war. Als Alternative wäre es besser, die lexikonbasierten Morphologieprogramme zu verwenden. In diesem Beitrag wurde entschieden kein Stemming durchzuführen, um herauszufinden, welche Ergebnisse damit zu erhalten sind.

Ein zusätzliches Problem bei der Datenverarbeitung hat sich aus der Mailpräsentation ergeben. In den Spam-Emails wurden höchst informative Wörter versteckt, wobei sie durch andere Zeichen zerlegt oder vertikal geschrieben wurden, wie es in der Abbildung 1 verdeutlicht wurde.

A
n
g
e
b
o
t

Abbildung 1. Das Problem der Spam-Präsentation.

Sind die Emails automatisch behandeln, ist diese Information verloren.

Während der Vorverarbeitung der Nicht-Spam-Nachrichten wird die Information über die Newsgroups entfernt, da sie die Klassifikation einfacher machen würde.

4.2 Testsettings

Die Mails werden als Ketten von Atributtenwerten dargestellt. Im Falle des Rainbow-Werkzeugs werden die Meldungen in einzelnen Dateien gespeichert, für Timbl werden sie aber zusammen abgespeichert. Zur Darstellung der Daten für Timbl verwendet man Fbinary Format, weil alle Attribute binär sind.

Als Lernalgorithmus wird IB1 und mit dem Ähnlichkeitsmaß'overlap metric'. Die Attribute werden durch Einsatz von 'Information Gain weighting' gewichtet.

Als Validation ist One-leave-out zu benutzen, die eine Möglichkeit bietet, jede Instanz und den Fehler des Klassifikators in zuverlässiger Weise zu testen. Beim Einsatz des Verfahrens des nächsten Nachbarn wird One-leave-out als auch 10-fold cross Validation verwendet. In diesem Fall werden die Daten in 10 Mengen unterteilt und der Lernalgotithmus zehnmal angewandt, um jede Menge zu testen (der Rest gilt dabei als Trainingdaten).

4.3 Ergebnisse

Da die Resultate der früheren Forschungen zeigen, dass die Leistungen des ähnlichkeitsbasierten und des Bayes-Verfahrens fast gleich sind, wurden in diesem Fall die besten Ergebnisse mittels des naïve-Bayes-Verfahrens erzielt (siehe Tab.1 und Tab.2).

Die Effektivität der eingesetzten Methoden lässt sich durch die Vollständigkeit (Recall) und Genauigkeit (Precision) messen, die die Anzahl der korrekt klassifizierten Emails im Verhältnis zur Anzahl aller korrekten und zur Anzahl aller klassifizierten Emails widerspiegelt.

5 Diskussion

Das Einsetzen von ähnlichkeitsbasiertem Lernen und der naïve-Bayes-Methode bietet eine gute Möglichkeit, die

[2] Eine Ausnahme bilden nur die erotischen Spam-Meldungen

Verfahren	Mails	Recall	Precision
Naïve Bayes	Spam	98,59%	99,76%
	Nicht-Spam	99,20%	95,38%
	Precision	98,73%	
Das Verfahren des nächsten Nachbarn	Spam	69,60%	98,86%
	Nicht-Spam	99,76%	91,77%
	Precision	92,91%	

Tabelle 1: Die Ergebnisse (One-leave-out)

Verfahren	Mails	Recall	Precision
Das Verfahren des nächsten Nachbarn	Spam	66,67%	88,89%
	Nicht-Spam	97,60%	91,30%
	Precision	90,91%	

Tabelle 2: Die Ergebnisse (10-fold Cross Validation)

Spam-Nachrichten zu filtern. Die in der Tab.1 gezeigten Werte sind außerdem höher als erwartet. Wie bereits erwähnt, wurde kein Stemming-Werkzeug verwendet, was zu redundanten Attributen führte. Man hatte diese Attribute nicht reduziert, um einzuschätzen, wie diese zwei Methoden sie behandeln.

Obwohl das ähnlichkeitsbasierte Verfahren schlechtere Resultate gebracht hat, ist zu betonen, dass es die hohe Genauigkeit und Vollständigkeit für die Nicht-Spam-Nachrichten liefert. Wenn man diese Ergebnisse mit solchen von [Hedstroem *et al.*, 2003] vergleicht, lässt sich zeigen, dass die letzten nicht viel besser sind (die Vollständigkeit für Spam-Meldung betrug 74,71%). Es ist aber zu verdeutlichen, dass man früher nicht die gleichen Daten verwendet hat, und auch nur ähnlichkeitsbasiertes Lernen benutzt wurde).

Wenn man die gleiche unerwünschte Email schon früher erhalten hat, werden die neuen Mails einfacher als Spam gekennzeichnet. Die erhaltenen Zahlen lassen sich dadurch erklären, dass die Meldungen kurz waren, also hat der Fakt, dass kein Stemming-Werkzeug verwendet wurde, keinen großen Einfluss auf die Klassifikation.

Einige Forscher berichten, dass in dem Gebiet der Spam-Erkennung eine sinnvolle Anzahl an Emails benutzt werden soll. Hierbei geht es darum, dass die neue Meldung aufgrund einer Trainingsmenge klassifiziert werden muss, die so klein wie möglich ist. Wenn man aber ein beispielbasiertes Verfahren benutzt, entsteht das Problem der ungesehenen Daten, u. zw., solche neue Muster, die zu keinen Daten aus Trainingmenge ähnlich sind, werden höchst wahrscheinlich falsch klassifiziert.

Die Untersuchung bestätigt auch die Hypothese, dass die zu denselben Themenbereichen gehörenden Spam- und Nicht-Spam-Emails sich schwieriger klassifizieren lassen. Fast alle falschen Voraussagen betrafen solche Daten.

Im wesentlichen sind die Ergebnisse befriedigend. Man sollte aber weitere Evaluationen durchführen, um herauszufinden, in welchen Fällen beide Verfahren nicht ausreichend sind. Man könnte die Ergebnisse durch die morphologische Analyse sowie durch das Hinzufügen der strukturellen Information verbessern.

Danksagung

Ich bedanke mich bei Kurt Spanier (Zentrum für Datenverarbeitung, Universität Tübingen) und Ulli Horlacher für die zur Verfügung gestellten Daten.

Literatur

[Androutsopoulos *et al.*, 2000] Androutsopoulos, I., G. Paliouras, V. Karkaletsis. *Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and A Memory-Based Approach*. Proceedings of the Workshop Machine Learning and Textual Information Access", European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), pp. 1-13, Lyon, France, 2000.

[Crawford *et al.*, 2001] Crawford E., J. Kay, E. McCreath. *Automatic Induction of Rules for e-mail Classification*. Proceedings of the Sixth Australasian Document Computing Symposium, Coffs Harbour, Australia, December 7, 2001.

[Daelemans *et al.*, 2002] Daelemans W., J. Zavrel, K. van der Sloot, and Antal van den Bosch. *TiMBL: Tilburg Memory Based Learner, version 4.2, Reference Guide*. ILK Technical Report 02-01, Available from http://ilk.kub.nl/downloads/pub/papers/ilk0201.ps.gz, 2002.

[Hedstroem *et al.*, 2003] Hedström A., S. Katrenko, Ch. Larsson. *Filtering Spam Mail Using Memory-Based Learning (TiMBL)*. Presentations in Information Retrieval. Students Workshop in Tübingen, May 29th - 31th, 2003.

[Marquez, 2000] Llus Marquez. Machine Learning and Natural Language Processing. Technical Report LSI-00-45-R, Departament de Llenguatges i Sistemes Inform0tics, Universitat Polit8cnica de Catalunya, 2000.

[Manco *et al.*, 2002] Manco G., E. Masciari, M. Ruffolo, A. Tagarelli. *Towards An Adaptive Mail Classifier*. In Italian Association for Artificial Intelligence Workshop Su Apprendimento Automatico: Metodi Ed Applicazioni, 2002.

[McCallum, 1996] McCallum, A. Kachites. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow. 1996.

[Pantel and Lin, 1998] Pantel P. and D. Lin. SpamCop: A Spam Classification & Organization Program. In Proceedings of AAAI-98 Workshop on Learning for Text Categorization. pp. 95-98. Madison, Wisconsin, 1998.

[Provost, 1999] J. Provost. Naive-Bayes vs. Rule-Learning in Classification of Email. University of Texas at Austin, Artificial Intelligence Lab. Technical Report AI-TR-99-284.