DATA-MINING-CUP 2003 – Aufgabe, Ergebnisse und Praxisschlüsse

Jens Scholz
prudsys AG
D-09113, Chemnitz, Deutschland
scholz@prudsys.de

DATA MINING CUP 2003

Um speziell Hochschulabsolventen an das Thema Data Mining und Analytical CRM heranzuführen, wurde im Jahr 2000 der DATA-MINING-CUP (DMC) Wettbewerb ins Leben gerufen, welcher seitdem alljährlich unter Beteiligung namhafter Hochschulen, Partner und Sponsoren stattfindet.

Nähere Informationen zum DMC sind unter http://www.data-mining-cup.de zu finden.

1 DMC 2003 Aufgabe

Die Aufgabe zum DMC 2003 orientierte sich an dem aktuellen und praktischen Problem der Erkennung von E-Mail-Spam.

Kernstück eines Mail-Filters ist seine Fähigkeit, E-Mails auf Grund verschiedener erhobener Merkmale zu klassifizieren und mit einer entsprechend hohen Wahrscheinlichkeit die Zuordnung Spam oder Ham (Nicht-Spam) zu treffen.

Eine weit verbreitete Software zur Spam-Filterung ist das Tool SpamAssassin. Der SpamAssassin beschreibt eine E-Mail mittels ca. 830 Merkmale und ermittelt die Klassenzugehörigkeit im Regelfall über eine einfache lineare Regression. Diese Diskrepanz zwischen komplexer Erhebung der Merkmale und vergleichsweise einfachem Klassifikationsverfahren ist ebenso bei vielen anderen Filtertools anzutreffen.

Beim Einsatz des SpamAssassin oder anderer Filtersoftware ergeben sich somit folgende Probleme:
- Qualität der Score-Berechnung, d.h. Trennschärfe i.d.R. unzureichend
- Anpassung des Klassifikationsziels eher schwierig

Lösung könnte hier ein eigener Klassifikator sein, der basierend auf den Merkmalsinformationen z.B. des SpamAssassin erstellt wird.

Der Wettbewerb zum DMC 2003 sah diese Aufgabenstellung vor.

Im Rahmen des Data Mining Cups waren dazu exemplarisch Daten von 8.000 E-Mails, sowie deren Klassenzugehörigkeit vorgegeben.

Aufgabenstellung war, einen Klassifikator zu erstellen, welcher in der Lage ist Ham-Mails von Spam-Mails zu unterscheiden. Im Rahmen des DMC war der Klassifikator dann exemplarisch auf 11.177 E-Mails anzuwenden und das Klassifikationsergebnis einzureichen.

2 DMC 2003 Ergebnisse

Die Einreichungen zum DMC 2003 bestätigten die Erwartungen, weit bessere Trennschärfen zu erreichen (bis zu 20-fach), als mit den Standardverfahren der Filter-Tools.

Der Gewinner des DMC 2003 (TU Ilmenau) konnte auf die Filterung von 99,5 Prozent der Spam-Mails verweisen. Wichtig war hierbei aber auch der geringe Anteil von nur 0,6 Prozent falsch-positiven E-Mails, d.h. Ham-Mail die fälschlicherweise als Spam erkannt wird.

Nach einer Befragung unter den Wettbewerbsteilnehmern zeigte sich, dass bei dieser Aufgabenstellung insbesondere nicht-lineare Verfahren Vorteile bei der Klassifikationsgüte erreichten. Sehr oft war unter den vorderen Plätzen somit unterschiedliche Implementierungen der SVM (Support Vektor Maschine) und Nicht-Lineare-Entscheidungsbäume zu finden. Als Herausforderung zeigte sich die Arbeit mit den verhältnismäßig großen Merkmalsvektoren des SpamAssassin (833 Attribute). Eine Dimensionsreduktion wurde somit von fast allen Teilnehmern, meist in Form einer Faktorenanalyse, durchgeführt.

3 Praxisschlüsse

Die prudsys AG (www.prudsys.de) hat parallel zur Aufgabe des DMC 2003 den Ansatz zur Verbesserung der E-Mail-Klassifikationsgüte in einer Data-Mining-Komponente umgesetzt. Die SpamClassifier gennante Komponente basiert auf der als GNU-Lizenz für den nicht-kommerziellen Einsatz auch frei verfügbaren XELOPES Data-Mining-Bibliothek (www.xelopes.de). Der SpamClassifier verbessert die E-Mail Klassifikation des SpamAssassin erheblich und kann auch eigene Klassifikatoren via PMML Schnittstelle (XML basiertes Austauschformat für Prediktive Modelle) einladen. Die Erstellung von Klassifikatoren ist damit vollständig Entkoppelt und kann mit jedem PMML-exportfähigem Data-Mining-Tool durchgeführt werden.