Rheinisch-Westfälische Technische Hochschule Aachen
Lehrstuhl für Informatik VI
Prof. Dr.-Ing. Hermann Ney

Praktikum Data Mining Cup im SS 2005

Praktikumsprotokoll

Betreuer: Thomas Deselaers, Daniel Keysers

Abstract

Der Data Mining Cup ist ein jährlicher Wettbewerb der Firma Prudsys, welcher auf internationaler Ebene Universitäten und Interessierte dazu einlädt, eine praxisnahe Aufgabe zu bearbeiten und gegebenenfalls auf den Datamining-Anwendertagen zu präsentieren. Der Lehrstuhl für Informatik 6 an der **RWTH** veranstaltete im Sommersemester 2005 ein Praktikum, in dessen Rahmen am DMC 2005 teilgenommen wurde.

1 Einleitung

Data Mining ist ein sehr praxisnaher Anwendungsbereich in der Informatik. Für viele Zwecke ist es nützlich oder sogar notwendig, in beliebigen Datenmengen nicht explizit enthaltene Informationen zu finden, wie zum Beispiel im Krankenversicherungsbereich, wo aus allgemeinen Kundendaten Informationen über die Wahrscheinlichkeit eines Versicherungsfalls gewonnen werden können. Der generelle Ansatz ist, mit Daten in denen die gewünschte Implikation bekannt ist einen Algorithmus zu trainieren, welcher dann in der Lage ist neue Datensätze zu klassifizieren. Nicht immer ist hierbei die niedrigste Fehlerrate signifikantes Optimierungskriterium, oft werden allen möglichen (Miss-)Klassifikationen Kosten zugeteilt, welche die Berechnung eines "Scores" ermöglichen. Desweiteren ist eine sehr gute Vorverarbeitung der Daten nötig, welche z.B. durch Ersetzung fehlender Werte oder Normalisierung einen starken Einfluss auf die Performance der Klassifikations-Algorithmen haben.
In diesem Protokoll soll die Vorbereitung auf die Aufgabe des Data Mining Cups 2005, so wie die Umsetzung des Gelernten vorgestellt werden.

2 Einarbeitungsphase

Zu Beginn der Einarbeitungsphase wurden die Grundlagen des Data Minings durch die Betreuer vorgestellt. Dabei wurde ein besonderes Augenmerk auf die Vermeidung des sogenannten **overfittings** gelegt, das "Übertrainieren" eines Algorithmus auf den Trainingsdaten. Weiterhin wurden die grundlegenden Methoden zur Datenvorverarbeitungen, die wesentlichen Klassifikationsalgorithmen, sowie einige Softwarelösungen für Data Mining vorgestellt.

Im weiteren Verlauf der Einarbeitungsphases sollten von den Praktikumsteilnehmern die grundlegenden Klassifikatoren Nearest Neighbour, Gauss und Naive Bayes selbst implementiert und auf Datensätzen älterer Data-Mining Wettbewerbe getestet werden. Es wurde jedoch bald klar, dass es für ein erfolgreiches Teilnehmen am Data Mining Cup 2005 wichtiger ist, den Umgang mit professioneller Software zu lernen, daher wurde die Abgabe der Programme verschoben und ein neuer Schwerpunkt gesetzt. Für viele Teilnehmer erwies sich das Programm Weka[1] wegen der Vielzahl an enthaltenen Klassifikatoren als das geeignetste Tool zum Testen von Vorverarbeitungen und Algorithmen. Eine weitere umfangreiche Sammlung stellt das Paket Netlab[2] dar. Weka ist durch seine GUI besonders gut geeignet, um an einem Datensatz verschiedene Klassikiationsansätze zu testen, NETLAB hingegen bietet effizientere und aktuellere Algorithmen. Zusätzlich stand der Maximum Entropy Klassifikator vom **I6** zur Verfügung, wobei als Nebenprodukt einige Skripte/Programme entwickelt wurden um Datensätze in die jeweils benötigten Formate zu konvertieren.

3 Data Mining Cup 2005

Die Aufgabe des Data Mining Cup 2005 wurde wieder von der Firma **Prudsys**[3] gestellt, und bestand aus einer Vorhersage von Betrugswahrscheinlichkeiten von Kunden eines Online-Versandhauses. Die Datensätze bestanden hier aus sehr vielen binären Features, einigen nominellen features mit einigen sowie mit vielen Ausprägungen, sowie ein paar numerischen Features. Zu Anfang des Praktikums wurden die Daten intensiv untersucht und vorverarbeitet. Für diese grundlegenden Untersuchungen erwies sich (für mich) **Matlab**[4] als sehr gute Arbeitsumgebung, da man Operationen auf einzelnen oder mehreren Feature-Vektoren sehr einfach und sehr schnell implementieren und durchführen kann. Lediglich der Daten Import/Export ist (gerade bei nicht-numerischen Features) teilweise mühselig. Bevor verschiedene Vorverarbeitungen getestet werden konnten, wurde ein Teil des Trainingsdatensatzes als **Holdout-Set** beiseite gelegt, um ein zu starkes Overfitting zu Vermeiden. Dadurch war es in der letzten Praktikumswoche möglich, optimierte Datensätze/Klassifikatoren auf Overfitting zu überprüfen. Insgesamt bauten hier die Arbeitsgruppen aufeinander auf, in wöchentlichen Treffen wurden neue Ergebnisse(Scores) und Transformationen von Features vorgestellt. Im Folgenden werden einige interessante Transformationen gezeigt.

3.1 Artikelnummern

Die Nummern der Artikel, die ein Kunde bereits bestellt hat, lagen in zehn Features gespeichert vor, wodurch bei Kunden die weniger als zehn Artikel bestellten viele Missing Values auftraten. Eine Untersuchung der Artikelnummern ergab, dass einige Artikelgruppen, eingeteilt nach der ersten Ziffer der Artikelnummer, stark erhöhte Betrugswahrscheinlichkeiten aufwiesen (siehe Figure 1). Hier konnten binäre oder numerische Features generiert werden, welche die Information ob, oder wieviele, Artikel einer Gruppe der Kunde gekauft hat, repräsentieren.

3.2 Kombinationsfeatures

Da, besonders nach einer grundlegenden Vorverarbeitung, sehr viele binäre Features vorhanden waren, war es einfach einen Zusammenhang zwischen Features mit bestimmten Wert und der Zielvariable zu untersuchen. Dadurch konnten neue Features

[1] http://www.cs.waikato.ac.nz/ml/weka/
[2] http://www.ncrg.aston.ac.uk/netlab/
[3] http://www.prudsys.de
[4] http://www.mathworks.com

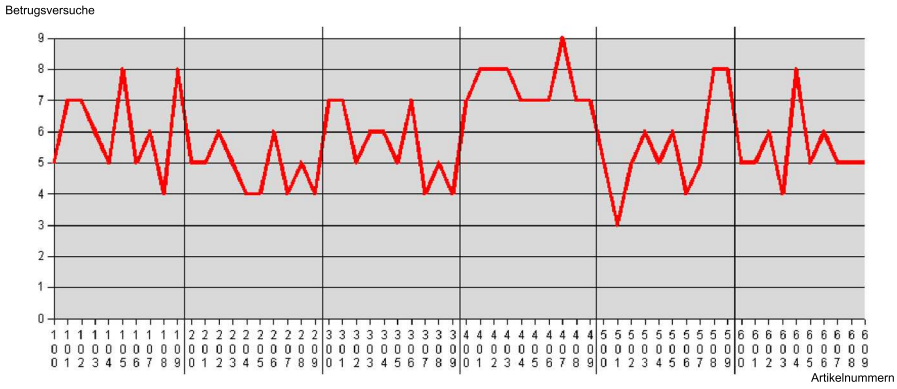

Figure 1: Betrugsversuche pro Artikelnummergruppe

wie z.B. *nicht-Neukunde AND nicht-Email_angegeben* erstellt werden. Um hier Overfittung zu vermeiden musste geprüft werden, ob die aus dem Trainingsset gewonnenen Informationen mit den Crossvalidation-sets[5] konsistent waren.

Zu Beginn wurde viel mit dem **Naive-Bayes** Klassifikator gearbeitet, da dieser sehr schnell arbeitet und sich leicht verbessern liess. Bald wurde aber deutlich, dass auf gegebenem Datensatz und mit den meisten Vorverarbeitungen **logistische Regression** die besten Ergebnisse erzielt, welches daher zum Baseline-Klassifikator wurde. Zum Abschluss des Praktikums wurde ein Tag gemeinsam an den möglichen Abgabe-Lösungen gearbeitet und diskutiert. Dadurch konnten einige Scores verbessert werden.

4 Abgegebene Lösungen

Die Lösungen unserer Gruppe wurden beide mit dem kostensensitiven Metaklassifikator mit logistischer Regression erstellt. Die zugrundeliegenden Datensätze bestanden aus normalisierten numerischen features, vielen binären features wie Artikelgruppierungen, Bestellzeitdiskretisierung und einer Auswahl an Kombinationsfeatures.

5 Fazit

Das Praktikum war sehr lehrreich und interessant. Durch die wöchentlichen Treffen konnte man sehr gut auf Ergebnissen der anderen Gruppen aufbauen, bzw. selber Grundlagen für neue Experimente liefern. Insgesamt war hohe Eigeninitiative gefordert, um sich nötiges Wissen und Fähigkeiten anzueignen. Hoffentlich tragen diese dazu bei, dass ein Teilnehmer des DMC-Praktikums einen vorderen Platz im DMC 2005 belegen oder sogar gewinnen kann.

[5] [n-fold]Crossvalidation ist ebenfalls eine Technik zur Vermeidung von Overfitting. Dabei wird das Trainingsset in n Testsets unterteilt, und für jedes Testset ein Trainingsset aus den gesamten Daten minus dem Testset erstellt.