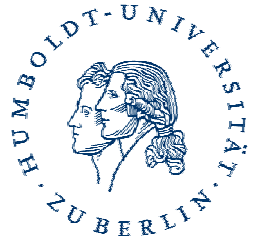

Exposé zur Studienarbeit

Maschinelles Lernen zur Schadensvorhersage bei Naturkatastrophen

Franziska Brosy

12. April 2008

Betreuer: Prof. Dr. Ulf Leser, Dr. Heidi Kreibich (GFZ)

Hintergrund

Die Sektion Ingenieurhydrologie am GeoForschungsZentrum Potsdam (GFZ) als Bereich des Departments Geoengineering beschäftigt sich unter anderem mit dem Risikomanagement hydrologischer Extreme. Ein Teilprojekt ist das Erfassen von Hochwasserschäden und dessen schadensbestimmenden Parametern [GFZ pb54, 2008]. Hierbei werden Kenntnisse über die Zusammenhänge zwischen den durch eine Flut entstandenen Schäden und deren beeinflussenden Faktoren gewonnen. Darauf aufbauend werden Methoden zur Schadensabschätzung entwickelt mit dem Ziel daraus optimale Strategien zur Schadensreduzierung ableiten zu können [GFZ pb54, 2008].

Ein Jahr nach dem großen Hochwasser an Elbe und Donau im August 2002 wurden Daten zu diesem Flutereignis mittels einer Telefonumfrage erhoben [Thieken et. al, 2007]. Zudem sind Daten zu den Hochwassern in der Stadt Dresden, die sich in den Jahren 2005 und 2006 ereigneten vorhanden [Kreibich et. al, 2008]. Diese Daten stammen ebenfalls aus Telefoninterviews. Für die Befragungen wurden zufällig Privathaushalte als auch Unternehmen aus den betroffenen Gebieten

ausgewählt. Mittels eines Fragebogens bestehend aus 180 Fragen zu den Schäden am Gebäude und am Hausrat, zu Fluteigenschaften, zu Frühwarnsystemen und zu vielen weiteren Faktoren, die den Flutschaden beeinflusst haben könnten, wurden Haushalte aus Einfamilien- und Mehrfamilienhäusern befragt. Die so entstandenen Datensätze wurden bereits statistisch ausgewertet [Thieken et. al, 2007] sowie einer Regressionsanalyse [Thieken et. al, 2005] unterzogen. Doch wurden hierbei nicht alle Flut-beeinflussenden Faktoren betrachtet, da die Datensätze recht komplex sind. Demzufolge besteht ein Bedarf der Anwendung anderer Analysemethoden [Thieken et. al, 2005].

Ziele

In einer Kooperation zwischen der Sektion Ingenieurhydrologie des GFZ und dem Institut für Informatik der Humboldt-Universität zu Berlin soll eine Analyse der Daten mit Methoden des maschinellen Lernens erfolgen. Dabei soll ermittelt werden, ob mit Hilfe von Klassifikationsalgorithmen neues verwendbares Wissen generierbar ist. Des Weiteren soll geklärt werden, welche Faktoren für eine möglichst präzise Schadensvorhersage besonders diskriminativ/wichtig sind. Bei der Datenanalyse soll nach den vier verschiedenen Hochwassertypen (langsame Flussüberschwemmung, Sturzflut, Deichbruch und Grundhochwasser [Hristova, 2007]) unterschieden werden.

Die resultierende Schadensschätzung als Schätzung des Gesamtschadens (Schaden am Gebäude und Schaden am Hausrat) wird in zweierlei Notationen angegeben: Zum einen in Form eines diskreten Wertes, der die Schadenshöhe in € wiedergibt, und zum anderen als relativer Wert, der den Schädigungsgrad in % beschreibt.

Abschließend soll eine Evaluation der ermittelten Hypothesen pro Lernalgorithmus erfolgen.

Daten

Anhand der Daten zu den Hochwassern im Elbe und Donau Gebiet in den Jahren 2003, 2005 und 2006 soll die Zweckdienlichkeit der Anwendung von Methoden des maschinellen Lernens zur Schadensvorhersage bei Naturkatastrophen evaluiert werden. Die Verwendung der Daten von unterschiedlichen Hochwasserereignissen hat den Vorteil, dass sich eine größere Verallgemeinerung der Vorhersagen erwarten lässt, als würden nur die Schadensdaten eines Extremereignisses in die Analyse einbezogen werden.

Die insgesamt 2158 Datensätze sind folgendermaßen auf die drei Flutereignisse verteilt:

Anzahl Datensätze	Hochwassergebiet	Jahr
1697	Elbe und Donau	2002
305	Dresden	2005
156	Dresden	2006

Ein Datensatz umfasst jeweils rund 1200 Spalten. Die Anzahl der Spalten spiegelt jedoch nicht die Anzahl der zu untersuchenden Merkmale wieder. Zusätzlich zu den Merkmalen sind viele weitere Informationen in den Datensätzen enthalten. Hier ist gegebenenfalls ist eine Aufbereitung der Daten vorzunehmen. Insofern, dass diejenigen Spalten automatisiert oder manuell entfernt werden, die nicht zur Repräsentation des Merkmalsraumes beitragen.

Vorgehen

Die Daten liegen zur Ansicht im SPSS-Datenformat vor und werden für die weitere Bearbeitung in ein Textformat (*.txt oder *.csv) exportiert. Die vorliegenden Datensätze werden in der Studienarbeit ohne weitere manuelle Bearbeitung benutzt.

Zur Realisierung einer möglichst präzisen Schadensvorhersage wird gegebenenfalls eine Merkmalsselektion vorgenommen, so dass die wichtigen Merkmale herausgestellt werden und ein stabiles Lernmodell gebildet werden kann. Um bei der Schadensvorhersage das Verhalten der vier verschiedenen Fluttypen betrachten zu können, werden die Datensätze in vier Partitionen geteilt und folglich getrennt analysiert.

Für die Analyse der Daten werden drei Klassifikationsmethoden des maschinellen Lernens angewendet. Dabei sollen Methoden aus den folgenden Bereichen zum Einsatz kommen:

- ein Algorithmus zum Entscheidungsbaum-Lernen,
- Bayes’sches Lernen (mit dem naive Bayes classifier) und
- Lernen linearer Klassifikatoren mit einer Kernel-Maschine (Support Vector Machine).

Es ist nicht Gegenstand der Studienarbeit genannte Ansätze eigens zu implementieren. Stattdessen soll eine Software ausgesucht werden, die bereits diese Algorithmen bereitstellt, um sie direkt auf den Daten anwenden zu können. Gegebenenfalls müssen die Daten in ein Software-spezifisch lesbares Format konvertiert werden.

Abschließend soll eine Auswertung erfolgen, in der aufgezeigt wird, welche Methode geeignet ist, um auf den gegebenen Daten eine möglichst präzise Schadensvorhersage zu treffen. Dazu werden je Klassifikationsmethode die Hypothesen evaluiert. Durch das Berechnen eines empirischen Risikos wird ein Schätzer für das echte Risiko angegeben (Verfahren: Holdout-Testing oder Cross Validation). Zudem werden jeweils die Entscheidungsfunktionen bewertet, indem je ein Qualitätsmaß ermittelt wird (wie ROC-Analyse oder Precision-Recall-Kurven). Zur schriftlichen Ergebnisdarstellung zählt auch die Beschreibung der Charakteristika der Algorithmen.

Literatur

[Thieken et. al, 2005] A. H. Thieken, M. Müller, H. Kreibich and B. Merz: Flood damage and influencing factors: New insights from the August 2002 flood in Germany, *Water Resources Research, 41, 12, W12430*, 2005

[Thieken et. al, 2007] A. H. Thieken, H. Kreibich, M. Müller and B. Merz: Coping with floods: preparedness, response and recovery of flood-affected residents in Germany 2002, *Hydrological Sciences Journal - Journal des Sciences Hydrologiques*, 52, 5, 1016-1037, October 2007

[Kreibich et. al, 2008] H. Kreibich, A. H. Thieken: Coping with floods in the city of Dresden, Germany, *Natural Hazards*, February 2008

[GFZ pb54, 2008] Ingenieurhydrologie - Projekte: Erfassung von Hochwasserschäden und schadensbestimmenden Parametern beim Hochwasser von Elbe und Donau im August 2002, *verfügbar: http://www.gfz-potsdam.de/pb5/pb54/projects/DamageSurvey/content.html*, (Januar, 2008)

[Hristova, 2007] B. Hristova: Analysis of the difference of flood damages caused by riverine flood, flash floods, levee breaches, and rising groundwater, *Master Thesis at Brandenburgische Technische Universität Cottbus*, August 2007