



Humboldt-Universität zu Berlin
Institut für Informatik
Fachgebiet Wissensmanagement in der Bioinformatik

Maschinelles Lernen zur Schadensvorhersage bei Naturkatastrophen

Studienarbeit

eingereicht von: **Franziska Brosy**

Betreuer: Prof. Dr. Ulf Leser

Dr. Heidi Kreibich (GFZ)

November 2008

Inhaltsverzeichnis

1	Einleitung.....	1
1.1	Motivation.....	1
1.2	Zielstellung	2
2	Grundlagen.....	3
2.1	Feature Selection.....	3
2.2	Klassifikationsalgorithmen	5
2.2.1	Entscheidungsbaum-Lernen mit C4.5	5
2.2.2	Bayes'sches Lernen mit dem Naiv Bayes Klassifikator.....	6
2.2.3	Lernen mit einer Support Vector Machine.....	8
2.3	Kurzeinführung zum Umgang mit dem Tool Weka.....	9
2.3.1	Das Laden der Daten.....	10
2.3.2	Das Ausführen und Anwenden der Feature Selection.....	10
2.3.3	Klassifikation	12
3	Realisierung.....	13
3.1	Definition des Lernproblems	13
3.2	Datenbasis.....	14
3.3	Datenaufbereitung mit SPSS und Perl	16
3.3.1	Automatisiertes Bilden der Schadensklassen und alphabetische Spaltensortierung in SPSS.....	16
3.3.2	Automatisiertes Zerlegen der Daten in je einen Datensatz pro Hochwassertyp.....	17
3.3.3	Automatisiertes Erzeugen von ARFF-Dateien	18
3.4	Schadensvorhersage.....	19
3.4.1	Daten laden und vorverarbeiten	20
3.4.2	Klassifikationsalgorithmen anwenden	21
3.4.3	Performanz der Klassifikation messen	22
4	Auswertung.....	24
4.1	Fragestellungen.....	24
4.1.1	Wie gut sind die Vorhersagen auf den Hochwasserdaten im Durchschnitt?	25

4.1.2	Gibt es einen Algorithmus unter den drei angewendeten, der eine bessere Vorhersage liefert als die anderen?	26
4.1.3	Führt eine zuvor angewendete Feature Selection zu einer Verbesserung des Klassifikationsergebnisses?.....	26
4.1.4	Welche Eigenschaften der Hochwasserereignisse sind besonders diskriminativ? .	28
4.1.5	Gibt es Unterschiede zwischen den Hochwassertypen oder zwischen den Schadenstypen bzgl. der Güte der Vorhersage?	33
4.1.6	Wird der Schaden generell eher unter- oder überschätzt?	35
4.1.7	Wie verändern sich die Ergebnisse bei einer separaten Untersuchung von nur nominalen bzw. nur numerischen Attributen?.....	36
5	Zusammenfassung und Ausblick	38
	Literatur- und Quellenverzeichnis.....	40
	Anhang	42
A	Ergebnisauswertung.....	42
A-1	Klassifikationsergebnisdiagramme	42
A-2	Classified-as Matrizen.....	44

Abbildungsverzeichnis

Abbildung 2-1	Bayes Hypothese gemäß [Scheffer, 2006].....	7
Abbildung 2-2	Problem der Mehrdeutigkeit von Trennebenen und Lösung mit der SVM.	8
Abbildung 2-3	Das Laden der Daten im Format ARFF in den Weka Explorer.....	10
Abbildung 2-4	Das Ausführen der Feature Selection mit dem Weka Explorer.....	11
Abbildung 2-5	Das Anwenden der Feature Selection innerhalb des Weka Explorers.....	11
Abbildung 2-6	Auswählen der Klassifikationsmethode innerhalb des Weka Explorers.	12
Abbildung 3-1	Evaluierungsergebnis am Beispiel von hw_1_result_gclass5_DT.	23
Abbildung 4-1	Klassifikationsergebnisse für den Hochwassertyp 1 und beide Schadenstypen. .	33
Abbildung 4-2	Klassifikationsergebnisse für den Hochwassertyp 3 und 4, je für den Gebäudeschaden.	34
Abbildung 4-3	Vergleich der Klassifikationsergebnisse für den Hochwassertyp 1 nominal versus numerisch.	37

Tabellenverzeichnis

Tabelle 2.1	Vor- und Nachteile beim Entscheidungsbaum-Lernen	6
Tabelle 2.2	Vor- und Nachteile beim Lernen mit dem Naiv Bayes Klassifikator	7
Tabelle 2.3	Vor- und Nachteile beim Lernen mit einer SVM.....	9
Tabelle 3.1	Verteilungen des Gesamtdatensatzes auf die Datensätze je Hochwassertyp	15
Tabelle 4.1	Hochwassertypenkürzel und ihre Bedeutung	24
Tabelle 4.2	Schadenstypenkürzel und ihre Bedeutung	25
Tabelle 4.3	Vergleich der Algorithmen bzgl. der Klassifikationsergebnisse je Kombination aus Hochwassertyp und Schadenstyp.....	26
Tabelle 4.4	Anzahl der korrekt klassifizierten Instanzen mit und ohne Feature Selection	28
Tabelle 4.5	Ergebnisse der Feature Selection je Kombination aus Hochwassertyp und Schadenstyp	32
Tabelle 4.7	Anzahl unter- und überschätzter Instanzen für alle Daten je Schadenstyp und Klassifikationsmethode.....	35

1 Einleitung

1.1 Motivation

Die Sektion Ingenieurhydrologie am GeoForschungsZentrum Potsdam (GFZ) als Bereich des Departments Geoen지니어ing 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 Strategien zur Schadensreduzierung ableiten zu können.

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]. Für die Befragungen wurden zufällig Privathaushalte 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 flutbeeinflussenden Faktoren betrachtet, da die Datensätze recht komplex sind. Demzufolge besteht ein Bedarf der Anwendung anderer Analysemethoden [Thieken et. al, 2005].

Zu den Hochwasserereignissen im Raum Dresden in den Jahren 2005 und 2006 wurden ebenfalls Daten per Telefonumfragen erfasst. Eine Betrachtung unterschiedlicher Hochwasserereignisse für die Analyse mit den Klassifikationsmethoden hätte eine bessere Verallgemeinerung der Vorhersagen bedeutet. Die Überarbeitung dieser Daten, so dass ein einheitlicher Datensatz (bezüglich der zu untersuchenden Eigenschaften) mit den Daten vom Hochwasser aus dem Jahr 2002 entstanden wäre, erwies sich jedoch als zu erheblicher Zeitaufwand. Diese zusätzlichen Daten wurden somit in dieser Arbeit nicht in die Analyse einbezogen.

1.2 Zielstellung

Anhand der Daten zum Hochwasser im Elbe und Donau Gebiet im Jahre 2002 soll in einer Kooperation zwischen der Sektion Ingenieurhydrologie des GFZ Potsdam und dem Institut für Informatik der Humboldt-Universität zu Berlin eine Analyse dieser Daten mittels drei unterschiedlichen Klassifikationsmethoden erfolgen. Dabei sollen Methoden aus den folgenden Bereichen zum Einsatz kommen:

- ein Algorithmus zum Entscheidungsbaum-Lernen,
- Bayes'sches Lernen (mit dem Naiv Bayes classifier) und
- Klassifikationslernen mit einer Kernel-Maschine (Support Vector Machine).

Ziel der Untersuchung ist es, herauszufinden, ob mit der Hilfe von Klassifikationsalgorithmen neues verwendbares Wissen aus den Daten generierbar ist. Des Weiteren soll geklärt werden, welche Faktoren für eine möglichst präzise Schadensvorhersage besonders diskriminativ sind. Bei der Datenanalyse soll nach den vier verschiedenen Hochwassertypen (langsame Flussüberschwemmung, Sturzflut, Deichbruch und Grundhochwasser [Hristova, 2007]) unterschieden werden. Pro Hochwassertyp soll eine Schätzung des Schadens zum einen für den Schaden am Gebäude und zum anderen für den Schaden am Hausrat erfolgen. Der vorhergesagte Schaden soll in Form einer diskreten Klasse, die die Schadenshöhe in € wiedergibt, repräsentiert werden. Abschließend soll eine Evaluierung der ermittelten Vorhersagen pro Lernalgorithmus erfolgen.

2 Grundlagen

Gemäß [Mitchell, 1997] beschäftigt sich das Feld des maschinellen Lernens mit der Frage, wie man Computer-Programme so konstruiert, dass diese ihre eigene Performanz (im Sinne von Qualität) verbessern können und zwar basierend auf Erfahrungswerten. Es wurden zahlreiche Programme entwickelt, die in der ersten Phase anhand von Beispielen (Daten mit bekannter Zielklasse) ein Modell trainieren. Das heißt, es werden die Eigenschaften der Daten genutzt, um Gesetzmäßigkeiten zu erkennen. Basierend auf den Gesetzmäßigkeiten erfolgt dann eine Zuordnung eines jeweiligen Datensatzes zu einer Zielklasse. In einer zweiten Phase wird getestet ob die Zuordnung zu einer Zielklasse mit der tatsächlichen Zielklasse übereinstimmt. Ist die Gesamtbilanz der Übereinstimmungen schlechter als eine angegebene Toleranzschwelle, wird versucht das Modell dahingehend zu verändern, dass sich die Trefferquote erhöht, also das Vorhersageergebnis verbessert. Dies geschieht über das Neubewerten einzelner Gesetzmäßigkeiten und das Neuberechnen der Zielfunktion.

Für die Untersuchung, ob aus den Hochwasserdaten mit Methoden des maschinellen Lernens neues Wissen generierbar ist, wurden in dieser Arbeit Ansätze aus drei Kernbereichen des maschinellen Lernens zur Vorhersage diskreter Werte ausgewählt. In diesem Kapitel werden die drei verwendeten Klassifikationsmethoden vorgestellt. Zuvor erfolgt eine Einführung zum Ansatz der Feature Selection.

2.1 Feature Selection

Der Begriff Feature Selection steht für den Prozess des Auswählens einer Teilmenge von Merkmalen (synonym auch Attribute/Eigenschaften), die für eine bestimmte Problemstellung von Bedeutung sind. Das heißt, es werden die Merkmale selektiert, die sich in Hinblick auf die gegebene Zielklasse besonders diskriminativ verhalten.

Das Anwenden einer Feature Selection ist dann sinnvoll, wenn die Daten mehrere Hundert oder Tausende von Merkmalen aufweisen [Yu and Liu, 2004]. Denn bei einer hohen Anzahl an Merkmalen wird sehr wahrscheinlich eine große Teilmenge davon, bezogen auf das Klassenkonzept, keinen informativen Beitrag leisten. Die Reduktion des Merkmalraumes durch das Entfernen von redundanten oder irrelevanten Features aus den Daten bewirkt eine Erhöhung der

Rechengeschwindigkeit der Lernalgorithmen, eine Verbesserung der Genauigkeit des Klassifikators und eine Vereinfachung des Lernmodells [Last, 2008]. Das Ermitteln einer Feature Selection teilt sich gemäß [Liu and Motoda, 1998] in die nachstehenden Teilfragen auf:

- Wie können die besten Features gefunden werden?

Die Feature Selection wird als ein Suchproblem angesehen, bei dem es gilt eine optimale Teilmenge (Subset) zu finden und zwar mittels Algorithmen, die den Merkmalsraum netzartig traversieren. Gibt es bereits generierte Subsets, dann werden Bäume an Stelle von Netzstrukturen durchsucht.

- Wodurch ist ein „bestes“ Feature charakterisiert?

Mögliche Bewertungskriterien werden sorgfältig untersucht und ihre Charakteristika werden analysiert. Unterschiedliche Ähnlichkeitsmaße werden dann entsprechend ihrer speziellen Anwendung eingesetzt.

- Wie sollen neue Features für die Auswahl behandelt werden?

Die bereits vorhandene Merkmalsauswahl wird überprüft, indem die Features bewertet werden. Anschließend wird das beste Merkmal aus der nicht ausgewählten Merkmalsmenge zu der Menge der ausgewählten Merkmale hinzugefügt oder es wird das unwichtigste Merkmal aus der Menge der ausgewählten Merkmale gelöscht. Bei der Betrachtung von Merkmal-Subsets an Stelle eines einzelnen Merkmals wird nach dem besten Subset gesucht.

In dieser Arbeit wurde der Algorithmus der „Correlation-based Feature Selection“ (CfsSubset, gemäß [Hall, 1998]) zur Evaluierung der Merkmale angewendet und als Suchmethode ist die Methode „BestFirst“ zum Einsatz gekommen. Demnach beruht die Feature Selection auf Korrelationen. Genauer gesagt realisiert die angewendete Java-Klasse CfsSubsetEval.java des Tools Weka (Kapitel 2.3) die Berechnung einer Pearson Korrelation der aktuellen Werte mit den Klassen, wobei die Varianz durch Entropie-Terme definiert wird (konditionale Entropie). Es werden also Unsicherheits-Koeffizienten berechnet und damit eine Korrelations-Matrix gefüllt. Kontinuierliche Attribute werden zuvor in diskrete Attribute konvertiert ([Hall and Smith, 1997] und [Hall and Smith, 1998]). Die BestFirst-Suchmethode der Java-Klasse BestFirst.java von Weka durchsucht den Merkmalsraum nach Merkmalsubsets mittels eines greedy hillclimbing Algorithmus mit zusätzlicher backtracking Möglichkeit [Weka, 2008].

2.2 Klassifikationsalgorithmen

Alle Klassifikationsmethoden haben die gleichen Voraussetzungen und ein einheitliches Ziel. Die Eingabe für die Klassifikationsalgorithmen bilden Merkmalsvektoren¹ und jeweils ein Zielattribut (eine Klasse) [Scheffer, 2006]. Zusammen repräsentieren diese beiden Komponenten die Trainingsdaten. Die Ausgabe beim Klassifikationslernen ist der Klassifikator² als eine Funktion, die Vektoren aus der Menge der Merkmalsvektoren auf eine Menge von Klassen abbildet.

2.2.1 Entscheidungsbaum-Lernen mit C4.5

Entscheidungsbaum-Lernverfahren klassifizieren Instanzen durch das Sortieren selbiger baumabwärts [Mitchell, 1997]. Das Ergebnis der Klassifikation (die Klassenwerte) stehen dann in den Blattknoten. Alle inneren Knoten sind Testknoten, die jeweils ein Merkmal repräsentieren. Hier wird getestet, ob der Wert eines Merkmals eine Bedingung erfüllt. Es erfolgt entsprechend des Testergebnisses die Verzweigung in einen Ast. Dies wird für jeden Teilbaum rekursiv weitergeführt.

Das Ziel des Lernens ist ein Modell, also einen Entscheidungsbaum zu finden, der zumindest für die Trainingsdaten möglichst die richtige Klasse liefert, wobei unter all den verschiedenen Bäumen, die mit den Trainingsdaten konsistent sind, der kleinste Baum ausgewählt werden soll [Scheffer, 2006]. Ein Grund ist, dass kleine Bäume sind leichter zu interpretieren und generalisieren meist besser. Letzteres bedeutet, dass sich die Klassenentscheidungen in den Blättern auf mehr Beispiele stützen können. Einige Vor- und Nachteile dieses Verfahrens sind in der Tabelle 2.1 gegenübergestellt.

Vorteile	Nachteile
Leicht zu Interpretieren. Liefern die Klassifikation (im Blatt) und zugleich die Begründung dafür (Pfad von der Wurzel zum Blatt).	Komplexität der vollständigen Suche nach einem Entscheidungsbaum ist exponentiell.

¹ Eine Menge von Datensätzen, die wiederum aus einer definierten Menge von Merkmalen pro Datensatz bestehen.

² Das gelernte Modell.

Einfacher Lernalgorithmus, der effizient und gut skalierbar (durch entsprechende Heuristiken) ist.	Das Verfahren ist anfällig für Overfitting, das durch starke Partitionierung entstehen kann.
--	--

Tabelle 2.1 Vor- und Nachteile beim Entscheidungsbaum-Lernen

Die Vorgehensweise des C4.5 als ein Algorithmus des Entscheidungsbaum-Lernens wird nachstehend skizziert. Die Implementierung, die auf den Hochwasser-Daten angewendet wurde, entspricht dem Algorithmus von [Quinlan, 1993].

Die Konstruktion eines Entscheidungsbaumes mit dem Algorithmus C4.5:

- Eingabe der Trainingsdaten
- Gibt es nur eine Klasse für alle zu untersuchenden Instanzen?
→ Ausgabe: Blatt mit dieser Klasse.
- Gibt es nur identische Instanzen in der Menge der Trainingsdaten?
→ Ausgabe: Blatt mit der häufigsten Klasse dieser Instanzen.
- Andernfalls suche den besten Testknoten, indem über alle Attribute iteriert wird und für jedes Attribut der Information Gain (siehe [Mitchell, 1997]), der das Splitkriterium repräsentiert, berechnet wird. Dabei unterscheidet sich die Berechnung für diskrete und kontinuierliche Attribute. Das Attribut mit dem höchsten Information Gain wird zum Testknoten.

Ist das aktuell zu untersuchende Attribut diskret?

- Teile die Trainingsdaten entsprechend der möglichen Werte für dieses Attribut.
Rekursion in alle Kindbäume

Ist das aktuell zu untersuchende Attribut kontinuierlich?

- Teile die Trainingsdaten in eine linke und eine rechte Untermenge. Dabei gilt: links landen all die Instanzen, deren Wert bezüglich des Wertes des Testknotens kleiner oder gleich ist. Der rechten Teilmenge werden all jene Instanzen zugeordnet, deren Wert bezüglich des Wertes des Testknotens größer oder gleich ist. Rekursion für den linken und den rechten Zweig mit dem C4.5 jeweils auf der zugeordneten Instanzenmenge.

2.2.2 Bayes'sches Lernen mit dem Naiv Bayes Klassifikator

Bei dem Naive-Bayes-Verfahren werden die Wahrscheinlichkeiten ermittelt, mit denen eine Instanz in einer Klasse vorkommt. Das bedeutet, der Bayes'sche Ansatz klassifiziert eine neue Instanz x_* .

mit ihren Merkmalswerten $(x_{*1}, x_{*2}, \dots, x_{*n})$, indem der Instanz die wahrscheinlichste Zielklasse y_* zugeordnet wird [Mitchell, 1997].

$$y_* = \arg \max_y \prod_{i=1}^n p(x_{*i} | y, \theta_{MAP}^{x_i|y}) p(y | \theta_{MAP}^y)$$

Naiv Bayes Klassifikator

Abbildung 2-1 Bayes Hypothese gemäß [Scheffer, 2006]. Die Berechnung des wahrscheinlichsten y minimiert das Risiko einer falschen Aussage. Die Annahme x_* hat n unabhängige Attribute repräsentiert den naiven Anteil und führt zum Aufmultiplizieren der einzelnen Wahrscheinlichkeiten, die die Aussagen, wie gut das gegebene Modell zu den Daten passt, liefern.

Auch dieses Verfahren besteht aus zwei Phasen. In der ersten Phase erfolgt der Lernvorgang. Das heißt, es werden die Modelle θ_{MAP}^y und $\theta_{MAP}^{x_i|y_j}$ aus den Trainingsdaten gelernt. Dabei spiegelt θ_{MAP}^y die relative Häufigkeit des Merkmals für die Klasse y in den Trainingsdaten wider und $\theta_{MAP}^{x_i|y_j}$ die relative Häufigkeit für das Merkmal i mit dem Wert x_i und Klasse y in den Trainingsdaten [Scheffer, 2006] angibt. Basierend auf dem Bayes'schen Theorem für bedingte Wahrscheinlichkeiten und mit der naiven Zusatzannahme, dass alle Attribute (Merkmale) paarweise voneinander unabhängig sind, ergibt sich der Naiv Bayes Klassifikator gemäß der Formel in Abbildung 2-1. Unter Verwendung der Modelle θ_{MAP}^y und $\theta_{MAP}^{x_i|y_j}$ wird für eine neue Instanz über den Naiv Bayes Klassifikator das Vorhersageergebnis berechnet. Einige Vor- und Nachteile dieses Verfahrens sind in der Tabelle 2.2 gegenübergestellt.

Vorteile	Nachteile
Funktioniert gut, selbst wenn die Unabhängigkeitsannahme nicht erfüllt ist.	Unabhängigkeitsannahme ist oft nicht erfüllt.
Einfach zu implementieren, effizient, populär. Einbezug von Anwendungswissen in Form von A-Priori-Wahrscheinlichkeiten wie $p(y \theta_{MAP}^y)$.	Schlechte Leistung wenn alle Merkmale bei mehreren Klassen etwa gleich verteilt sind.

Tabelle 2.2 Vor- und Nachteile beim Lernen mit dem Naiv Bayes Klassifikator

Die konkrete Implementierung des Naiv Bayes Klassifikators, die auf den Hochwasserdaten angewendet wurde, entspricht dem Algorithmus aus [John and Langley, 1995].

2.2.3 Lernen mit einer Support Vector Machine

Die binäre Klassifikation einer neuen Instanz mit einer Support Vector Machine (SVM) besteht aus dem Zuordnen der neuen Instanz zu einer Seite der im Lernvorgang ermittelten Trennebene. Bei der Klassifikation für ein 2-Klassenproblem wird die Seitenzuordnung der Trennebene bestimmt, indem der Wert $f(x)$ der Trennfunktion für die neue Instanz x berechnet wird. Der Wert für $f(x)$ ist entweder kleiner oder größer Null und dadurch ist bestimmt ob die vorhergesagte Klasse der linken oder der rechten Seite (mit der dort repräsentierten Klasse) der Trennebene entspricht.

Aus den Trainingsdaten eine Trennebene zu ermitteln, wenn die Instanzen linear separierbar sind, ist nicht immer schwer, aber auch nicht immer eindeutig - siehe Darstellung in der Abbildung 2-2. Support Vector Machines suchen nach einer Trennebene, die den Abstand (Margin) der Ebene von den Instanzen maximiert und dabei die Anzahl der den Margin verletzenden Instanzen (Slack-Terme) gering hält, um so die beste Trennebene für die Separation zu finden.

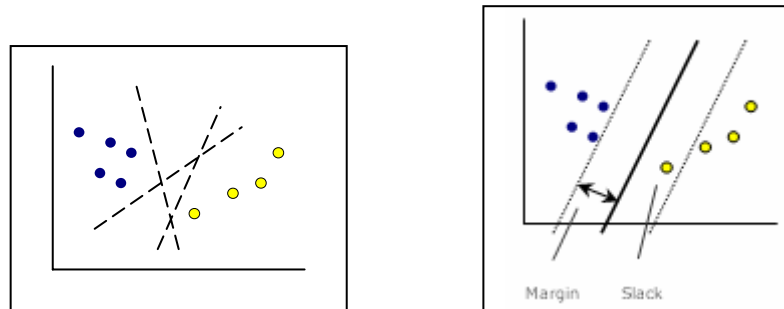


Abbildung 2-2 Problem der Mehrdeutigkeit von Trennebenen und Lösung mit der SVM. Oft lassen sich mehrere Trennebenen zwischen Instanzenmengen bspw. zweier verschiedener Klassen (Beispiele) finden. Die SVM findet genau eine Trennebene, die den Abstand zwischen der Ebene und allen Beispielen maximiert.

Sollten die Instanzen nicht linear separierbar sein, dann werden sie in einen anderen, höherdimensionalen Raum abgebildet, in dem sie wiederum linear klassifizierbar sind [Scheffer, 2006]. Damit Support Vector Machines auch für Multiklassen-Klassifikation (Klassifikation mit mehr als zwei Klassen) angewendet werden können, werden mehrere 2-Klassen-SVMs kombiniert, so dass beliebig viele Klassen unterschieden werden können [Eckstein, 2006]. Dazu gibt es zwei Ansätze. Ein Ansatz besteht in der Unterscheidung je einer Klasse vom Rest (alle anderen Klassen) und der andere Ansatz in der Unterscheidung von je zwei Klassen. Bei der Multiklassen-SVM bekommt die Trennfunktion einen zweiten Parameter y . Die vorhergesagte Klasse wird dann als $y^* = \arg \max_y f(x, y)$ [Scheffer, 2006] berechnet.

Einige Vor- und Nachteile dieses Verfahrens sind in der Tabelle 2.3 gegenübergestellt.

Vorteile	Nachteile
Kompakte Modelle und gute Klassifikatoren.	Die gefundenen Modelle sind schwer zu deuten. Unter Umständen gibt es lange Trainingszeiten.
Effiziente Klassifikation neuer Instanzen.	Aufwendige Implementierung.

Tabelle 2.3 Vor- und Nachteile beim Lernen mit einer SVM

Die konkrete Implementierung der Support Vector Machine, die auf den Hochwasser-Daten angewendet wurde, entspricht dem Algorithmus aus [Platt, 1998] und [Keerthi et. al, 2001].

2.3 Kurzeinführung zum Umgang mit dem Tool Weka

Das Tool Weka repräsentiert eine Sammlung von Algorithmen aus dem Bereich des maschinellen Lernens für Data Mining Aufgaben [Weka, 2008]. Die Algorithmen können direkt auf vorhandene Daten angewendet werden oder aus eigenem Java Code heraus aufgerufen werden. Weka enthält zudem Tools für das Vorverarbeiten von Daten, für Klassifikation, Regression, Clustering und Assoziationsregeln. Weka ist eine Open Source Software, die unter der GNU General Public License herausgegeben wird. In dieser Arbeit wurde der Weka Explorer von Weka in der Version 3.4.12 verwendet. Heruntergeladen werden kann das Tool über die Webseite <http://www.cs.waikato.ac.nz/ml/weka>. Wird Weka gestartet öffnet sich der Weka GUI Chooser, über den ein sogenannter Weka Explorer gestartet werden kann. Sollte sich der Weka Explorer bei der Anwendung von Lernalgorithmen auf größeren Datenmengen mit der Fehlermeldung „out of java heap space“ beenden, kann dieser Fehler durch das Verändern des Parameters „maxheap“ in der Datei RunWeka.ini behoben werden. Für die Anwendungen auf den Hochwasserdaten, die dieser Arbeit zu Grunde liegen, genügt es maxheap auf den Wert „1024m“ hochzusetzen.

Der Ablauf der Anwendung von Methoden des maschinellen Lernens auf den Daten innerhalb des Weka Explorers besteht im Wesentlichen aus drei Phasen, die in den Abschnitten 2.3.1 bis 2.3.3 beschrieben werden.

2.3.1 Das Laden der Daten

Die zu untersuchenden Daten, die gemäß des in Kapitel 3.3 beschriebenen Vorgehens erzeugt wurden, werden über „Open File“ auf dem Menütab „Preprocess“ in den Weka Explorer geladen. Der Vorgang des Ladens der Daten innerhalb von Weka über die graphische Benutzeroberfläche (GUI) ist in der Abbildung 2-3 skizziert.

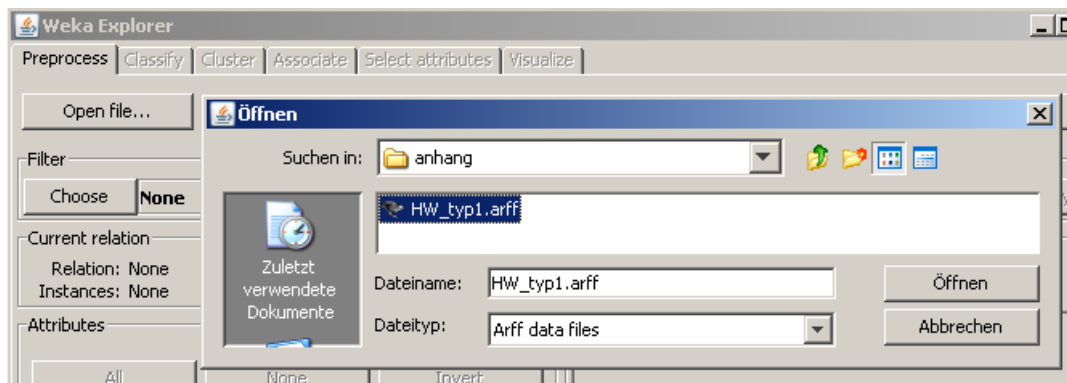


Abbildung 2-3 Das Laden der Daten im Format ARFF in den Weka Explorer

2.3.2 Das Ausführen und Anwenden der Feature Selection

Die Feature Selection ist in zwei Teilschritte unterteilt. Zuerst wird über das Menütab „Select attributes“ eine Auswahl der Merkmale berechnet, die für die angegebene Zielklasse besonders von Bedeutung ist (siehe Abbildung 2-4). Dazu kann aus einer Reihe in Weka angebotener Algorithmen sowohl eine Merkmalsevaluierung als auch eine Suchmethode ausgewählt und angewendet werden. Die Berechnung wird mittels des „Start“ Buttons angestoßen und als Ergebnis werden die Attributnamen der ermittelten Feature Selection aufgelistet.

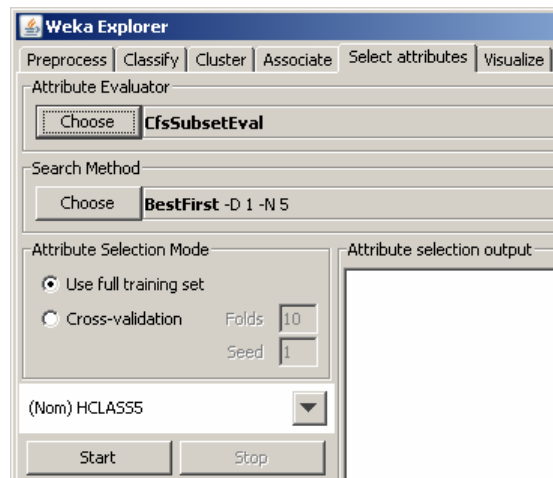


Abbildung 2-4 Das Ausführen der Feature Selection mit dem Weka Explorer. Die Feature Selection wird mit dem Algorithmus „CfsSubsetEval“ zur Evaluierung der Merkmale und mit der Suchmethode „BestFirst“ ausgeführt. Als Ergebnis wird eine Liste von Merkmalen geliefert, die für das Zielattribut besonders diskriminativ sind.

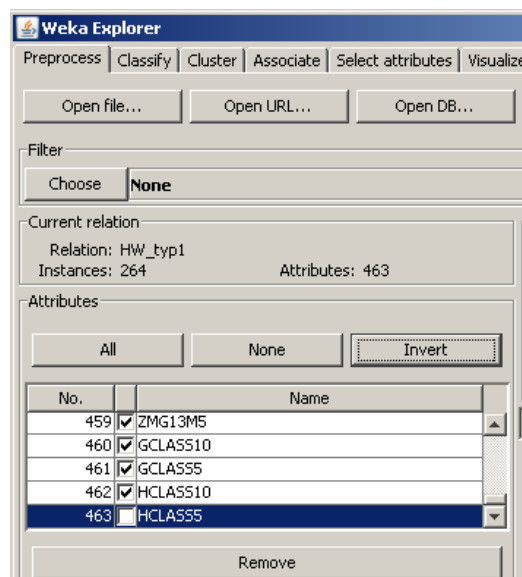


Abbildung 2-5 Das Anwenden der Feature Selection innerhalb des Weka Explorers. Das heißt, es werden aus der gesamten zur Verfügung stehenden Merkmalsmenge all jene Merkmale entfernt die nicht in der Feature Selection enthalten sind und die nicht das Zielattribut repräsentieren.

Anschließend kann das Ergebnis auf die zu untersuchenden Daten angewendet werden (siehe Abbildung 2-5). Dies geschieht indem im Bereich „Attributes“ auf dem „Preprocess“ Menütab all jene Attribute markiert (Häkchen setzen) und gelöscht („Remove“ Button) werden, die nicht zur berechneten Feature Selection gehören. Bei einem Datensatz mit sehr vielen Attributen ist es

effizienter alle Attribute zu markieren („All“ Button), dann für die Attribute aus der Feature Selection und für das Zielattribut (die Klasse) die Markierung wieder aufzuheben (Häkchen wegnehmen). Danach wird die Auswahl invertiert („Invert“ Button) und folglich werden alle nun markierten Attribute gelöscht („Remove“ Button).

2.3.3 Klassifikation

Auf dem „Classify“ Menütab lassen sich über den „Choose“ Button die Klassifikationsalgorithmen auswählen (siehe Abbildung 2-6). In dieser Arbeit wurden drei der zur Verfügung stehenden Algorithmen angewendet. Diese sind classifiers/trees/J48 für das Entscheidungsbaum-Lernen, classifiers/bayes/ NaiveBayes für das Bayes'sche Lernen mit dem Naiv Bayes Klassifikator und classifiers/ functions/SMO für das Lernen mit einer Support Vector Machine. Zusätzlich kann über die „Test options“ eine n-fold Cross Validation in die Klassifikation mit einbezogen werden. Nach der Auswahl des Klassenattributs wird über den „Start“ Button der Lernalgorithmus auf den zuvor geladenen Daten ausgeführt.

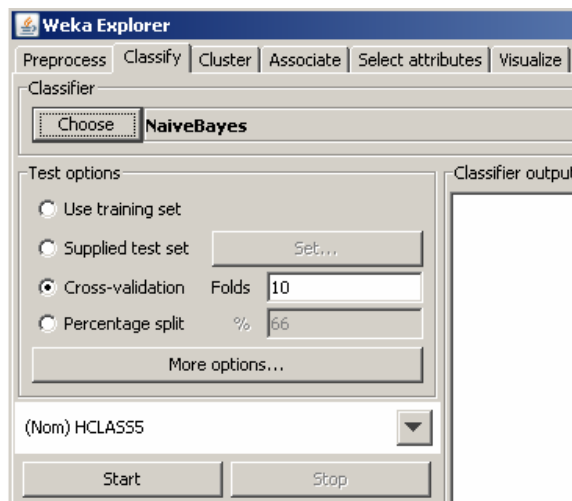


Abbildung 2-6 Auswählen der Klassifikationsmethode innerhalb des Weka Explorers. In dieser Abbildung wurde für die Klassifikation das Naiv Bayes Verfahren und eine 10-fache Kreuzvalidierung ausgewählt.

3 Realisierung

Der Weg von den in der Software SPSS vorliegenden Hochwasserdaten zu den Klassifikationsergebnissen beginnt mit der Definition des Lernproblems. Ist das Lernproblem klar formuliert, kann mit dem Tool Weka die Klassifikation stattfinden. Da für dieses Tool ein bestimmtes Dateneingabeformat Voraussetzung ist, muss zuvor eine Konvertierung der Daten in das für Weka spezifische Format erfolgen.

3.1 Definition des Lernproblems

Gemäß [Mitchell, 1997] ist das Lernproblem hinreichend gut gestellt, wenn die Aspekte Task, Performance measure und training Experience (kurz TPE) klar spezifiziert sind. Für die Schadensvorhersage bei Naturkatastrophen basierend auf den Hochwasserdaten ist das Prinzip TPE wie folgt definiert:

- T steht für die Klassifizierung neuer Schadensereignisse anhand der Eigenschaften der Schadensereignisse.
- P entspricht dem Ermitteln der prozentualen Angaben der richtig zugeordneten Schadensklassen zu den Schadensereignissen.
- E wird repräsentiert durch die Hochwasserdaten bestehend aus 601 Eigenschaften je Datensatz inklusive ihrer bekannten Schadensklassen.

Vervollständigen lässt sich das Lernsystem durch das Bestimmen der Zielfunktion als exakter Typ des Wissens, das gelernt werden soll sowie durch eine präzise Darstellung des Zielwissens und der Auswahl der Lernmethode [Mitchell, 1997].

- Die Zielfunktion ist die Abbildung $F: X \rightarrow C$ mit $X =$ Eigenschaftskonfiguration (oder auch Merkmalsbelegung) und $C =$ Schadensklasse.
- Die Zielfunktionsdarstellung eines linearen Klassifikators lautet dann:
$$\hat{f}(x) = w_0 + w_1x_1 + \dots + w_nx_n$$
 mit $n =$ die Anzahl der Schadensmerkmale und $w =$ die Gewichtung der einzelnen Merkmale.

- Die Lernmethode ist hier aufeinander folgend das Anwenden des Entscheidungsbaum-Lernens mit dem C4.5, des Bayes'schen Lernens mit dem Naiv Bayes Klassifikator und das Lernen mit einer Support Vector Machine.

3.2 Datenbasis

Die für diese Arbeit vorliegenden Daten zum Hochwasserereignis im Elbe und Donau Gebiet im August 2002 bestehen aus 1697 Datensätzen mit je 601 Merkmalen. Diese 601 Eigenschaften setzen sich aus verschiedenen Kategorien zusammen. Dazu gehört die Kategorie der Hochwassersituation, zu der unter anderem Informationen über den Wasserstand, die Fließgeschwindigkeit, die Art des Hochwassers und das Treibgut aufgenommen wurden. Weitere Kategorien sind die Gebäudesituation (Informationen über u.a. Art, Alter, Zustand, Etagen, Nutzung und Fläche des Gebäudes) und die Hausratsituation (Inventar, Heizsysteme, Bodenbeläge und Art der Räume: bspw. Sauna, Büro, Werkstatt, Wohnräume). Außerdem wurden Daten zur Kategorie der Wissenssituation der Bewohner erfasst, dazu zählen u.a. Kenntnisse über Vorsorgemaßnahmen, getroffene Vorsorgemaßnahmen, bereits miterlebte Hochwasserereignisse und die Kenntnisnahme von Hochwasserwarnungen.

Bei der Darstellung der Attribute wurde zwischen den zwei Datentypen nominal und numerisch unterschieden. Die Verteilung der Datentypen sieht wie folgt aus: von den 601 Attributen sind 58 vom Typ numerisch (reelle oder ganze Zahlen) und 543 vom Typ nominal (vordefinierte unsortierte Liste von Werten). Unter den nominalen Attributen befinden sich 5 Attribute, die echte Strings repräsentieren, während der Rest aus Aufzählungen von zumeist Integerwerten und zu einem geringen Teil aus Zahlen mit Nachkommastellen besteht. Von den nominalen Attributen mit Zahlwerten weisen 371 Attribute eine Listengröße von 1 auf, eine Listengröße von 2 oder 3 besitzen weitere 50 Attribute. Ein einzelnes Attribut weist die maximale Listengröße von 123 auf und die restlichen 116 nominalen Attribute haben durchschnittlich 20 Werte.

Informationen über den durch das Hochwasser entstandenen Schaden am Gebäude und am Hausrat liefern die Attribute `czgs18oz` (Gebäudeschaden in Euro) und `chs7aoz` (Hausratschaden in Euro). Unter den insgesamt 1697 Datensätzen befinden sich 618 Datensätze ohne Einträge für den Gebäudeschaden und 425 Datensätze sind ohne Einträge für den Hausratschaden.

Da es Gegenstand dieser Arbeit ist, die Schadensvorhersage für jeden der vier verschiedenen Hochwassertypen einzeln zu untersuchen, muss der gesamte Datensatz in vier einzelne Datensätze

zerlegt werden. Daraus ergeben sich die in der nachstehenden Tabelle 3.1 aufgeführten Verteilungen, des Gesamtdatensatzes auf die einzelnen Datensätze je Hochwassertyp:

Hochwassertyp	Anzahl der Instanzen je Hochwassertyp	Darunter befinden sich die folgenden Anzahlen von Instanzen, mit Angaben zum Gebäudeschaden ungleich NULL.	Darunter befinden sich die folgenden Anzahlen von Instanzen, mit Angaben zum Hausratschaden ungleich NULL.	Von den insgesamt 601 Attributen gibt es die folgende Anzahl an Attributen, die nicht komplett mit NULL Werten belegt sind.
Grundhochwasser	264	160	152	459
langsame Flussüberschwemmung	593	334	477	543
Sturzflut	578	385	417	526
Deichbruch	262	200	226	466

Tabelle 3.1 Verteilungen des Gesamtdatensatzes auf die Datensätze je Hochwassertyp

Anmerkung

In den Daten sind bei den nominalen String-Attributen Werte enthalten, die mit einer Ziffer beginnen. Strings müssen um sie später automatisiert als Strings erkennen zu können mit einem Buchstaben beginnen. Folglich wurden die drei Einträge

- ...,1 Biotonne, ...
- ...,2 riesige Baustellencontainer, ...
- ...,200 qm des Grundstückes waren vom Hochwasser weggerissen, ...

geändert in

- ...,eine Biotonne, ...
- ...,zwei riesige Baustellencontainer, ...
- ...,ca. 200 qm des Grundstückes waren vom Hochwasser weggerissen, ...

3.3 Datenaufbereitung mit SPSS und Perl

Für den Einsatz der Software Weka zum maschinellen Lernen auf den Daten vom Hochwasser an Elbe und Donau im August 2002 [Thieken et. al, 2007] müssen die als SPSS Datensatz vorliegenden Daten in das Weka spezifische Datenformat ARFF (Attribute-Relation File Format) umgewandelt werden. Eine ARFF-Datei ist eine ASCII-Textdatei, die eine Liste von Instanzen beschreibt und die ein bestimmtes Set an Attributen aufweist [Weka ARFF, 2008].

3.3.1 Automatisiertes Bilden der Schadensklassen und alphabetische Spaltensortierung in SPSS

Die Datenkonvertierung basiert auf der Umwandlung von einer Textdatei im Format (*.dat), in der die Daten Tabulator getrennt aus der Statistik Software SPSS exportiert wurden. Zuvor wurden die Originaldaten an die Anforderungen des maschinellen Lernens mit Klassifikationsmethoden angepasst. Das bedeutet, zu den Daten mit ihren 601 Attributen wurden Klassen (die Zielattribute) hinzugefügt. Zur Unterscheidung der zwei Schadenstypen Gebäudeschaden und Hausratschaden wurde jeweils eine Klasse eingefügt. Zusätzlich soll je Schadenstyp eine Unterteilung in fünf Klassen und eine Unterteilung in elf Klassen untersucht werden. Diese beiden Skalierungen sollen später Aufschluss darüber geben, ob feiner skaliert werden kann, bei gleich bleibender oder verbesserter Vorhersagegenauigkeit als bei gröberer Skalierung der Klassen. Demzufolge wurden an die Datensätze im SPSS vier Spalten angehängt, die anhand von Berechnungsformeln³ mit Daten basierend auf den Spalten czgs18oz⁴ für die Klassenbildung zum Gebäudeschaden bzw. chs7aoz⁵ für die Klassenbildung zum Hausratschaden befüllt wurden.

Eine weitere Vorbereitung der Daten für die spätere Anwendung des Tools Weka bestand in der alphabetischen Sortierung⁶ der Spaltennamen der nun 605 Spalten. Der Zweck eine

³ SPS-Skript zu Syntax Klassenbildung in SPSS befindet sich auf der der Arbeit beigelegten CD.

⁴ Gebäudeschaden in Euro (ohne den höchsten Wert).

⁵ Hausratschaden in Euro.

⁶ SPS-Skript zur Sortierung von Spaltennamen in SPSS befindet sich auf der der Arbeit beigelegten CD.

alphabetische Reihenfolge in den Spaltennamen zu erhalten, liegt im gezielten und zeiteffizienten Auffinden und Selektieren der Spalten innerhalb von Weka, die durch eine Feature Selection bestimmt wurden. Aufgrund einer sortierten Auflistung der Spaltennamen kann direkt zu einem Buchstabenbereich gescrollt werden, anstatt top-down oder bottom-up Spaltenname für Spaltenname zu betrachten. Jedoch gehen bei der Sortierung, die durch das Transponieren, anschließender Zeilensortierung und erneutem Transponieren der Daten realisiert wird, die Werte für String-Spalten verloren. Die Ursache hierfür ist, dass beim Transponieren im SPSS automatisch alle Spalten in numerische Spalten mit maximal 40 Stellen umgewandelt werden. Demzufolge mussten die Daten der Spalten

- o_xs1 (ohne Beschreibung im SPSS-Variablenlabel),
- mue_t (Mündungsbereich, betroffene Gewässer),
- o_e8a (ohne Beschreibung im SPSS-Variablenlabel),
- gkzo3_5 (ohne Beschreibung im SPSS-Variablenlabel),
- ort (Ort), plz (Postleitzahl), stra (Strasse)
- und der vier Klassen (gclass5, gclass11, hclass5, hclass11)

manuell nachgetragen werden, indem die Daten aus der Originaldatei kopiert und wieder in diese Spalten eingefügt wurden.

3.3.2 Automatisiertes Zerlegen der Daten in je einen Datensatz pro Hochwassertyp

Die weitere Aufbereitung der Daten wurde mit Hilfe von selbst geschriebenen Programmmodulen in der Sprache Perl automatisiert umgesetzt. Als Voraussetzung für das Anwenden der im Folgenden beschriebenen Perlskripte bedarf es der Installation von ActiveState Perl (www.activestate.com/store/activeperl/download).

Ausgeführt werden die Perlskripte dann aus der Kommandozeile heraus mit dem Befehl perl. Ein Perlskript `Zerlegung_in_4_HW_Typen.pl`⁷ realisiert die Zerlegung der gesamten Daten aus einer DAT-Datei in vier einzelne DAT-Dateien, wobei je Hochwassertyp eine Datei mit den Datensätzen, die zu dem entsprechenden Hochwassertyp gehören, erzeugt werden. In der ersten

⁷ Perlskript zur Datenzerlegung in 4 Hochwassertypen befindet sich auf der der Arbeit beigelegten CD.

Zeile einer solchen Datei stehen die Spaltennamen und die restlichen Zeilen repräsentieren die eigentlichen Daten. Die Unterscheidung des Hochwassertyps erfolgt anhand der Einträge in der Spalte hw_typ4⁸. Die Zerlegung ist notwendig, da die vier Hochwassertypen getrennt voneinander untersucht werden sollen.

Eingabe	aus_spss_aufbereitete_daten.dat
Aufruf	perl Zerlegung_in_4_HW_Typen.pl aus_spss_aufbereitete_daten.dat (vorausgesetzt das Skript und die Daten liegen im selben Verzeichnis; Aufruf von Perl aus diesem Verzeichnis)
Ausgabe	HW_typ1.dat, HW_typ2.dat, HW_typ3.dat und HW_typ4.dat

3.3.3 Automatisiertes Erzeugen von ARFF-Dateien

Ein weiteres Skript ARFF_aus_DAT_erstellen.pl⁹ erzeugt aus den einzelnen HW_typX.dat Dateien eine ARFF-Datei. Dieses Skript ist universell einsetzbar, sobald die folgenden Anforderungen an die DAT-Datei erfüllt sind:

- die einzelnen Spalten sind jeweils Tabulator getrennt,
- in der ersten Zeile stehen die Spaltennamen,
- in der zweiten Zeile stehen die Angaben zur Unterscheidung von nominalen und numerischen Spalten (Eintrag 1111 für nominal und 2222 für numerisch),
- um echte Strings sind doppelte Anführungszeichen gesetzt und
- Stringspalten beginnen nicht mit einem Zahlwert (wie bspw. „1 Biotonne“).

Die ARFF-Datei wird generiert, indem die nachstehenden Änderungen an den eingegebenen DAT-Dateien vorgenommen werden:

- das Löschen aller Punktzeichen,
- das Ersetzen aller Kommata durch Punktzeichen,

⁸ Einteilung der Interviews in 4 Hochwassertypen aufgrund von Flussabschnitten und Flussmorphologie.

⁹ Perlskript zur ARFF-Erzeugung aus einer DAT-Datei befindet sich auf der der Arbeit beigelegten CD.

- das Ersetzen aller Tabulatoren durch Kommata,
- das Ersetzen von .Zahl mit 0.Zahl und -.Zahl mit -0.Zahl,
- das Trimmen von Zelleneinträgen, indem führende und nachstehende Leerzeichen und Tabulatoren gelöscht werden,
- das Ersetzen fehlender Zelleneinträge mit einem Fragezeichen und
- das Setzen von doppelten Anführungszeichen um echte Strings.

Dieser Prozess erzeugt eine Hilfsdatei mit dem Namen `eingabedateiname_converted.dat`. Weiterhin wird ein ARFF-Header (Syntax siehe [Weka ARFF, 2008]) für die Datenquelle erzeugt, indem aus den zwei Kopfzeilen die Metainformationen für die Spalten gefiltert werden. Das hat den Hintergrund, dass eine Beschreibung der Spalten in der Form:

- `@attribute spaltennameXyz numeric`
- `@attribute abc {4,1,2,3,5,6}` mit 1 bis 6 als mögliche Werte für die nominale Spalte `abc`

Bestandteil eines jeden ARFF-Headers ist. Das bedeutet, dass nominale Spalten von oben nach unten nach distinkten Werten durchsucht werden müssen, um die Werte zu der entsprechenden Aufzählungsliste hinzuzufügen. Da mit einer leeren Menge wie `@attribute xyz {}` eine ARFF-Datei nicht in das Weka Tool geladen werden kann, müssen diese leeren Spalten sowohl aus dem ARFF-Header als auch aus den eigentlichen Daten gelöscht werden. Dazu wird eine Hilfsdatei `help.dat` geschrieben, die die reinen Datensätze ohne Leerspalten enthält, um diese Daten im Abschnitt `@data` in der ARFF-Datei nach dem Erstellen des ARFF-Headers anzuhängen. Die Unterscheidung von nominalen und numerischen Attributen erfolgt anhand der Einträge in der zweiten Zeile in der eingegebenen DAT-Datei.

Eingabe	HW_typX.dat
Aufruf	<code>perl ARFF_aus_DAT_erstellen.pl HW_typX.dat</code>
Ausgabe	HW_typX.arff und die zwei Hilfsdateien (<code>help.dat</code> und <code>HW_typX_converted.dat</code>)

3.4 Schadensvorhersage

Basierend auf den Daten im ARFF-Format können nun über das Tool Weka die Modelle für die Schadensvorhersagen mit den drei verschiedenen Klassifikationsmethoden C4.5, Naiv Bayes und SVM gelernt werden. Dieser Trainingsprozess umfasst zwei Teilschritte. In Schritt 1 werden die Daten (*.arff Dateien) in das Tool Weka geladen und es wird eine Feature Selection auf den

eingeebenen Daten berechnet und angewendet. Im zweiten Schritt werden die Klassifikationsalgorithmen auf den Daten ausgeführt. Anschließend wird eine Performanzmessung für die einzelnen Klassifikatoren durchgeführt, um zu einer Aussage über die Güte der Klassifikation auf den Hochwasserdaten zu gelangen.

3.4.1 Daten laden und vorverarbeiten

Es liegen nach der Datenaufbereitung mit SPSS und Perl vier ARFF-Dateien (je Hochwassertyp eine) vor. In jeder dieser Dateien sind vier Klassenangaben enthalten, denn es gilt je Hochwassertyp ein Klassifikationsergebnis sowohl für den Hausratschaden als auch für den Gebäudeschaden zu ermitteln und es gilt es jeweils die Klasseneinteilung in fünf bzw. in elf Klassen zu untersuchen (siehe auch Kapitel 3.3.1). Da sich über die ersten Ergebnisse zum Hochwassertyp 1 herausgestellt hat, dass eine Unterteilung in elf Klassen zu ungenauen Klassifikationen liefert, wurde ab dem Hochwassertyp 2 nur noch die fünf Klassenvariante verwendet. Folglich wurde die Datei HW_typ1.arff viermal geladen, um für jede der vier Klassen die Feature Selection zu berechnen und anzuwenden. Für alle weiteren Dateien, das heißt für Hochwassertyp 2 bis 4 und ‚all‘¹⁰ erfolgte dieser Vorgang nur noch zweimalig, nämlich für die Klasseneinteilung in fünf Klassen je Schadenstyp. Insgesamt ergibt dies zwölf ausgeführte Ladevorgänge und Feature Selections.

Die Feature Selection wurde mit der Weka Standardkonfiguration CfsSubset als Evaluierer und BestFirst als Suchmethode berechnet (siehe Kapitel 2.1). Der Vorgang des Berechnens und Anwendens der Feature Selection innerhalb von Weka entspricht der Darstellung in Kapitel 2.3.2. Die Liste der jeweils selektierten Attribute ist der Tabelle 4.5 zu entnehmen.

Bemerkung 1

Bei der Feature Selection kann keine Spalte als Klasse (also als Zielattribut) ausgewählt werden, die vom Typ numerisch ist. Denn für Klassifikationsaufgaben sind Klassen immer nominal. Andernfalls würde es sich um Regressionsaufgaben handeln.

¹⁰ Gesamter Datensatz ohne die Zerlegung in die verschiedenen Hochwassertypen. Demzufolge sind hier alle vier Hochwassertypen vertreten.

Bemerkung 2

Die Feature Selection kann auch weglassen werden. Das Trainieren der Klassifikationsmodelle ist unabhängig von einer zuvor ausgeführten Feature Selection. Allerdings verbessert sich meist die Güte der Klassifikationsergebnisse durch den Einsatz der Feature Selection.

3.4.2 Klassifikationsalgorithmen anwenden

Das Anwenden der Klassifikationsalgorithmen mit einer 10-fold Cross Validation erfolgt wie in Kapitel 2.3.3 beschrieben innerhalb des Weka Explorers. Zusätzlich müssen noch die Ergebnisse der Klassifikation gespeichert werden. Gespeichert werden kann zum einen das trainierte Klassifikationsmodell und zum anderen eine ARFF-Datei mit den vorhergesagten Klassen für die Trainingsdaten.

Das Speichern der vorhergesagten Klassen geschieht über das Untermenü „Visualize classifier errors“, das über einen Rechtsklick mit der Maus auf den entsprechenden Lauf der Klassifikationsmethode im Feld „Result list“ erreichbar ist. Infolgedessen öffnet sich ein Fenster, auf dem sich ein „Save“ Button befindet, der es erlaubt das Ergebnis als ARFF-Datei zu sichern. In dieser Datei wird an die Trainingsdaten eine weitere Spalte (predictedKlassenname) angefügt, die die Vorhersageergebnisse beinhaltet. Ein beispielhafter Auszug aus einem ARFF-Header:

```
@attribute predictedHCLASS5 {1901-5000,0-1900,5001-11000,11001-25000}  
@attribute HCLASS5 {1901-5000,0-1900,5001-11000,11001-25000}
```

In den Daten spiegelt sich das Ergebnis in den entsprechenden Spalten am Beispiel wie folgt wieder:

- 11001-25000,11001-25000, wenn die Vorhersage für einen Datensatz mit der Klasse „11001-25000“ exakt getroffen wurde und
- 1901-5000,0-1900, wenn für einen Datensatz mit der Klasse „0-1900“ die Nachbarklasse „1901-5000“ vorhergesagt wurde.

Weiterhin lässt sich das gelernte Klassifikationsmodell aus der „Result list“ heraus speichern, indem der entsprechende Lauf markiert wird und durch einen Rechtsklick mit der Maus das Untermenü geöffnet wird, auf dem die Aktion „Save model“ ausgewählt werden kann. Infolgedessen kann das Modell als *.model Datei gespeichert werden. Sollen später für neue Instanzen ohne Klassenlabel mit Hilfe dieses gelernten Modells die Klassen vorhergesagt werden, kann für die Vorhersage das Modell geladen werden. Unter der Voraussetzung dass die neuen Daten ohne Klassenlabel in den Weka Explorer geladen wurden, kann über das „Classify“ Menütab mit einem Rechtsklick in das Feld der „Result list“ ein Untermenü geöffnet werden, über das wiederum „Load

model“ ausgewählt werden kann. Wird hier das gespeicherte Modell angegeben, kann über den „Start“ Button die Vorhersage der Klassenlabels basierend auf dem trainierten Modell erfolgen. Das Vorhersageergebnis neuer Instanzen lässt sich ebenso wie das Trainingsergebnis speichern.

In dieser Arbeit wurden je Ladevorgang gemäß der Beschreibung aus Kapitel 3.4.1 (Daten laden und vorverarbeiten) die drei Klassifikationsmethoden C4.5, Naiv Bayes und SVM ausgeführt. Somit wurden 36 Klassifikationsergebnisse (*_result.arff Dateien) erzeugt, für die nun noch eine Performanzmessung zu erfolgen hat.

Bemerkung 3

Anstatt der k-fold Cross Validation kann auch eine Leave-One-Out Cross Validation (LOOCV) beim Klassifizieren verwendet werden. Da die LOOCV einer k-fold Cross Validation mit $k = n = \text{Anzahl der Instanzen}$ entspricht [Scheffer, 2004], kann innerhalb von Weka für die Anzahl der Felder die Anzahl der Instanzen angegeben werden, um eine LOOCV zu erhalten. Das Verwenden einer LOOCV liefert eventuell verlässlichere Ergebnisse, erhöht aber wesentlich die Programmlaufzeit.

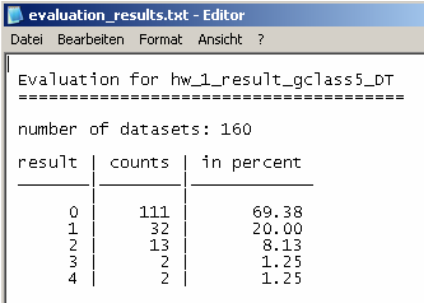
3.4.3 Performanz der Klassifikation messen

Nachdem die Modelle für das Klassifizieren neuer Schadensereignisse anhand der Trainingsdaten gelernt wurden, gilt es noch die Güte der Modelle zu bestimmen. Das heißt, es gilt zu klären: Wie viel Prozent der vorhergesagten Schadensklassen sind richtig zugeordnet? Oder auch wie viel Prozent richtig klassifizierter Schadensereignisse gibt es? Die Berechnung dieser Angaben erfolgt über das Perlskript `evaluierung.pl`¹¹.

Eingabe	HW_typX_result.arff
Aufruf	perl evaluierung.pl HW_typX_result.arff (Ergebnis an bereits generierte evaluation_results.txt anhängen; falls diese noch nicht existiert wird sie neu angelegt) perl evaluierung.pl HW_typX_result.arff clean (Ergebnis in zuvor geleerte evaluation_results.txt schreiben)
Ausgabe	evaluation_results.txt

¹¹ Perlskript zur Evaluierung der Schadensvorhersagen befindet sich auf der der Arbeit beigelegten CD.

Die Evaluierung besteht darin, die Anzahl der richtig zugeordneten Klassen zu zählen sowie auch die Anzahl der um eine Klasse verschätzten Klasse und der um zwei Klassen verschätzten Klasse und so weiter. Neben den absoluten Angaben werden die Angaben auch als prozentualer Anteil bezogen auf die eingegebene Datenmenge berechnet und ausgegeben.



```
evaluation_results.txt - Editor
Datei Bearbeiten Format Ansicht ?

Evaluation for hw_1_result_gclass5_DT
=====
number of datasets: 160
result | counts | in percent
-----|-----|-----
0      | 111    | 69.38
1      | 32     | 20.00
2      | 13     | 8.13
3      | 2      | 1.25
4      | 2      | 1.25
```

Abbildung 3-1 Evaluierungsergebnis am Beispiel von `hw_1_result_gclass5_DT`. Das Perlskript `evaluierung.pl` erzeugt für die als Argument übergebene ARFF-Datei, die das Vorhersageergebnis beinhaltet, eine Tabelle, die den Abstand der vorhergesagten Klasse zur echten Klasse („result“) und die Anzahl der Instanzen pro Klasse mit dem gegebenen Abstand („counts“ bzw. Angabe „in percent“) wiedergibt. Die Ausgabe des Skripts gemäß der Abbildung wird in eine Textdatei geschrieben.

Die Abbildung 3-1 repräsentiert das Evaluierungsergebnis des Entscheidungsbaumlernens (DT = Decision Tree) für die Kombination aus Hochwassertyp 1 und Gebäudeschaden. Die Anzahl der Datensätze, die in die Bewertung eingegangen sind beträgt 160. Die Spalte „result“ gibt an, wie groß der Abstand zwischen dem Vorhersageergebnis und der tatsächlichen Klasse in den Trainingsdaten ist. Wenn `result=0` ist, bedeutet das, dass die Instanz richtig klassifiziert wurde, `result=1` entspricht einer Vorhersage der Nachbarklasse etc. Jeweils drei Evaluierungsergebnisse (die Ergebnisse der drei untersuchten Klassifikationsmethoden pro Kombination aus Hochwassertyp und Schadenstyp) wurden in einem Diagramm zusammengefasst. Alle Diagramme sind im Anhang A-1 zu finden.

4 Auswertung

Die Klassifikationsergebnisse und die Performanzmessung zur Ermittlung der Güte der Vorhersagen, die die verschiedenen Klassifikationsmethoden auf den Trainingsdaten geliefert haben, bilden die Grundlage zur Klärung einiger Fragestellungen. Dabei steht im Fokus welche Unterschiede es in den Ergebnissen zwischen den einzelnen Klassifikationsmethoden gibt und welche Unterschiede zwischen den Hochwassertypen und auch zwischen den Schadenstypen hinsichtlich der Güte der Klassifikationsergebnisse bestehen.

4.1 Fragestellungen

Zunächst sollen die Abkürzungen, die in den Auswertungsdiagrammen- und tabellen Verwendung finden, erläutert werden. Die Bedeutung der Hochwassertypen 1 bis 4 und „all“ wird in der Tabelle 4.1 repräsentiert.

Hochwassertyp	Bedeutung
1	Grundhochwasser
2	langsame Flussüberschwemmung
3	Sturzflut
4	Deichbruch
all	Alle Daten, somit sind alle 4 Hochwassertypen enthalten.

Tabelle 4.1 Hochwassertypenkürzel und ihre Bedeutung

Die Klassen für die Bereiche der Schadensbeträge und die Unterscheidung der Schadenstypen anhand ihrer Kurzbezeichnungen sind in Tabelle 4.2 dargestellt.

Schadenstyp	Bedeutung	Bereiche der Schadensbeträge in €
gclass5	Schadensklassen für den Gebäudeschaden. Mit der Einteilung der Schadensbeträge in 5 Klassen.	0-5000; 5001-15000; 15001-30000; 30001-100000; 100001-755000

hclass5	Schadensklassen für den Hausratschaden. Mit der Einteilung der Schadensbeträge in 5 Klassen.	0-1900; 1901-5000; 5001-11000; 11001-25000; 25001-160000
---------	--	---

Tabelle 4.2 Schadenstypenkürzel und ihre Bedeutung

4.1.1 Wie gut sind die Vorhersagen auf den Hochwasserdaten im Durchschnitt?

Werden die korrekt klassifizierten Instanzen für alle Kombinationen aus Hochwassertyp und Schadenstyp betrachtet und werden die Ergebnisse aller Klassifikationsmethoden gemittelt, dann gilt, dass für die Hälfte aller Instanzen die Klassen richtig vorhergesagt wurden. Wird die Abweichung um eine Klasse nach oben oder nach unten in Betracht gezogen, so werden weitere 36% annähernd richtig vorhergesagt. Folglich werden für 86% aller Instanzen (mit einer Toleranz von einer Klasse nach oben oder nach unten) sehr gute Klassifikationsergebnisse gemittelt über alle verwendeten Klassifikationsmethoden erbracht. Die einzelnen Ergebnisse sind in der Tabelle 4.3 aufgeführt. Die Spalten „0“ stehen für die Anteile der richtig klassifizierten Instanzen und die Spalten „0+1“ stehen für die Anteile der richtig klassifizierten Instanzen plus die Anteile der als Nachbarklasse von der korrekten Klasse klassifizierten Instanzen.

	Bester Algorithmus		Entscheidungsbaum(DT)		Naiv Bayes (NB)		Support Vector Machine (SVM)	
	0	0+1	0	0+1	0	0+1	0	0+1
hw1gclass5	DT	NB	70,0	91,9	68,75	93,13	58,1	93,1
hw1hclass5	NB	NB	47,4	80,3	55,9	90,8	42,8	90,1
hw2gclass5	NB	NB	34,1	67,9	49,1	88,32	40,7	84,4
hw2hclass5	NB	NB	39,4	75,3	48,0	83,4	38,6	75,9
hw3gclass5	NB	NB	44,9	75,8	51,4	91,1	38,4	81,3
hw3hclass5	DT	DT	45,3	82,5	37,6	79,4	31,65	74,8
hw4gclass5	NB	SVM	65,5	91,5	67,0	91,5	62,5	92,0
hw4hclass5	DT	DT	61,9	90,7	51,8	90,3	52,6	87,6

hwAllgclass5	NB	NB	52,3	84,4	56,3	92,2	50,0	87,7
hwAllhclass5	DT	DT	52,0	87,4	48,0	86,6	43,5	84,1

Tabelle 4.3 Vergleich der Algorithmen bzgl. der Klassifikationsergebnisse je Kombination aus Hochwassertyp und Schadenstyp

Die Ergebnisse zu den Vorhersageabweichungen um 2 bis 4 Klassen von der richtigen Klasse sind den Diagrammen im Anhang A-1 zu entnehmen.

4.1.2 Gibt es einen Algorithmus unter den drei angewendeten, der eine bessere Vorhersage liefert als die anderen?

Aus der Tabelle 4.3 geht hervor, dass für die einzelnen Hochwassertypen insgesamt fünfmal der Naiv Bayes Klassifikator das beste Vorhersageergebnis bezüglich der korrekt klassifizierten Instanzen geliefert hat und auch fünfmal das beste Ergebnis für die Toleranz von einer Abweichung um eine Klasse von der richtigen Klasse. Hingegen dazu hat die Support Vector Machine nie als beste Methode bei den korrekt klassifizierten Instanzen abgeschlossen und nur einmal war die SVM besser als NB und DT in der Vorhersage mit der Toleranz von einer Klasse als Abweichung. Das Entscheidungsbaum-Lernen mit dem C4.5 liegt von der Klassifikationsleistung her zwischen dem NB und der SVM. Somit ist der Algorithmus mit den besten Vorhersageergebnissen auf den Hochwasserdaten bei Betrachtung der einzelnen Hochwassertypen der Naiv Bayes Klassifikator.

Aus der Tabelle 4.3 ist außerdem ersichtlich, dass gemittelt über alle Hochwassertypen (hwAll) in den beiden Bereichen (der richtig klassifizierten Instanzen und der richtig klassifizierten Instanzen inklusive der als Nachbarklasse klassifizierten Instanzen) der NB bei der Vorhersage des Gebäudeschadens die besten Ergebnisse erzielt, während es bei der Vorhersage des Hauratschadens der DT ist.

4.1.3 Führt eine zuvor angewendete Feature Selection zu einer Verbesserung des Klassifikationsergebnisses?

Das Trainieren auf einem ausgewählten Set von Attributen, das durch eine Feature Selection berechnet wurde, erzielt ein besseres Klassifikationsergebnis. So ist nach der Feature Selection die Anzahl der Instanzen mit korrekt vorhergesagten Klassen im Vergleich zur Klassifikation auf allen Attributen größtenteils, nämlich in 87,5% aller Berechnungen besser und für den geringen Rest von

12,5% annähernd gleich. Folglich trägt das Ausführen und Anwenden einer Feature Selection zur einer Verbesserung der Vorhersagen bei. Die Anzahl der korrekt klassifizierten Instanzen pro Algorithmus und Kombination aus Hochwassertyp und Schadenstyp sind in der nachstehenden Tabelle 4.4 zu finden.

		Anzahl richtig klassifizierter Instanzen ohne FS (601 Attribute)	Anzahl richtig klassifizierter Instanzen mit FS	Anzahl selektierter Attribute	Algorithmus
hw1gclass5	↑	110	112	8	DT
	↑	83	110	8	NB
	↑	90	93	8	SVM
hw1hclass5	↓	88	72	7	DT
	↑	45	85	7	NB
	↓	67	65	7	SVM
hw2gclass5	↑	112	114	11	DT
	↑	143	164	11	NB
	↑	103	136	11	SVM
hw2hclass5	↑	177	188	8	DT
	↑	182	229	8	NB
	↓	188	184	8	SVM
hw3gclass5	↑	160	173	9	DT
	↑	172	198	9	NB
	↑	146	148	9	SVM
hw3hclass5	↑	177	189	7	DT
	↑	134	157	7	NB
	↑	123	132	7	SVM
hw4gclass5	↑	128	131	16	DT
	↑	118	134	16	NB

	↑	120	125	16	SVM
hw4hclass5	↑	123	140	7	DT
	↑	104	117	7	NB
	↑	112	119	7	SVM

Tabelle 4.4 Anzahl der korrekt klassifizierten Instanzen mit und ohne Feature Selection

4.1.4 Welche Eigenschaften der Hochwasserereignisse sind besonders diskriminativ?

Aus der Tabelle 4.5, die die Ergebnisse der Feature Selection je Kombination aus Hochwassertyp und Schadenstyp beinhaltet, geht hervor, dass sich die Sets der diskriminativen Attribute sowohl abhängig vom Hochwassertyp als auch abhängig vom Schadenstyp unterscheiden. Zur Ermittlung der diskriminativen Attribute wurden in einem sukzessiven experimentellen Vorgehen diejenigen Attribute zuvor aus der Gesamtattributmenge entfernt, die aufgrund ihrer Bedeutung die Klassifikationsergebnisse verfälscht haben. So wurden bspw. die Klassen gclass10, hclass5 und hclass10 für die Klassifikation mit dem Zielattribut gclass5 aus der zu untersuchenden Attributmenge gelöscht (gclass5, gclass10 und hclass10 entsprechend für das Zielattribut hclass5). Diese Attribute sollten zwar inhaltlich keine Auswirkungen auf die Ergebnisse haben, jedoch haben Experimente gezeigt, dass auch diese Attribute von der Feature Selection ausgewählt werden und folglich auf das Vorhersageergebnis einen unerwünschten Einfluss haben. Gleiches gilt für die folgenden Attribute bei der Klassifikation des Gebäudeschadens:

cvsu_w	Korrektur: Wert2002 des Gebäudes: $w_g1914 * 13,1402$ [VSU in €] (Gleitender Neuwertfaktor 2002)
czgs18oz	Gebäudeschaden in Euro (ohne den höchsten Wert 4 Mio)
gco_xkoo	GCO_XKoord
gco_ykoo	GCO_YKoord
intid	Intid
ss_w	Berechnung: Schadensatz Wohngebäude ($czgs18oz / cvsu_w$)

Deshalb wurden auch diese Attribute aus der zu untersuchenden Attributmenge für den Gebäudeschaden gelöscht. Für die Vorhersage des Hausratschadens sind entsprechend die Attribute

chs7aoz	Hausratschaden in Euro.
cvsu_h2	vsu_h2 Korrektur: wenn chs7aoz>vsu_h2, dann cvsu_h2=chs7aoz, sonst cvsu_h2=vsu_h2
gco_xkoo	GCO_XKoord
gco_ykoo	GCO_YKoord
intid	Intid
ss_h2a	Schadensatz Hausrat (chs7aoz / cvsu_h2)

ausselektiert worden. Zusätzlich wurden für beide Schadenstypen die Attribute stra (Strasse), plz (Postleitzahl) und ort (Ort) entfernt, um Herauszufinden welche Attribute unabhängig von der Lage eines befragten Haushaltes eine besonders diskriminative Wirkung auf den entstandenen Schaden für diesen Haushalt haben. Gemeinsame Attribute, die aus den jeweiligen Feature Selections hervorgegangen sind

- für nahezu alle Hochwassertypen und nahezu alle Schadenstypen: ce1haupt (Korrekturspalte neu: Fluss, der hauptsächlich für den Schaden verantwortlich war.), chs11oz (Korrekturspalte: Höhe der schon erhaltenen finanziellen Entschädigung) und erf7 (Seit wann wohnen Sie\ Ihre Familie schon in der vom Augusthochwasser 2002 betroffenen Wohnung?),
- für alle Hochwassertypen und Gebäudeschaden: cgs20oz (Korrekturspalte: Teuerster Einzelschaden am Gebäude Wert in Euro (Zahl)), wstuegok (Berechnung: Wasserstand über GOK (Kellerwerte minus 2.5 m)) und zg6 (In welchem Jahr wurde das Gebäude erbaut?)
- und für alle Hochwassertypen und Hausratschaden: chs7oz (Korrekturspalte: Wiederbeschaffungswert teuerstes Objekt in Euro) und hs8 (Schätzen Sie bitte, wie viel Prozent Ihres Hausrates durch das Hochwasser zerstört wurde.).

Alle selektierten Attribute je Kombination aus Hochwassertyp und Schadenstyp sind in der Tabelle 4.5 aufgelistet. Die Angabe „#I“ gibt dabei die Anzahl der Instanzen wieder, die je Kombination zur Verfügung standen.

	#I	Selected attributes
HW Typ1 gclass5	160	<p><u>cgs20oz</u> - Korrekturspalte: Teuerster Einzelschaden am Gebäude Wert in Euro (Zahl)</p> <p><u>chs11oz</u> - Korrekturspalte: Höhe der schon erhaltenen finanziellen Entschädigung</p> <p><u>cs3</u> - Wohnen Sie zur Miete oder gehört Ihnen das Haus oder die Wohnung?</p> <p><u>erf7</u> - Seit wann wohnen Sie\ Ihre Familie schon in der vom Augusthochwasser 2002 betroffenen Wohnung?</p> <p><u>gs1z</u> - Zusammenführung der Mieter und Eigentümerangaben, liegen beide Antworten vor, gilt die des Eigentümers! (keine Angabe/weiß nicht gelöscht)</p> <p><u>wstuegok</u> - Berechnung: Wasserstand über GOK (Kellerwerte minus 2.5 m)</p> <p><u>zg6</u> - In welchem Jahr wurde das Gebäude erbaut?</p> <p><u>zgs16</u> - Wurden diese Arbeiten von Ihnen selbst durchgeführt, oder haben sie</p>

		Handwerker beauftragt? Schätzen sie bitte, wie viel Prozent der angefallenen Gesamt-Arbeitszeit von Handwerkern erbracht wurde.
HW Typ1 hclass5	152	<p><u>ce1haupt</u> - Korrekturspalte neu: Fluss, der hauptsächlich für den Schaden verantwortlich war.</p> <p><u>chs11oz</u> - Korrekturspalte: Höhe der schon erhaltenen finanziellen Entschädigung</p> <p><u>chs7oz</u> - Korrekturspalte: Wiederbeschaffungswert teuerstes Objekt in Euro</p> <p><u>e2</u> - Wie ist das Wasser in das Gebäude eingedrungen? Von außen, also z.B. durch Fenster, Türen, Keller- und Luftschächte oder von unten, also z.B. durch die Toilette, Waschbecken, Bodeneinläufe?</p> <p><u>erf7</u> - Seit wann wohnen Sie\ Ihre Familie schon in der vom Augusthochwasser 2002 betroffenen Wohnung?</p> <p><u>hs8</u> - Schätzen Sie bitte, wie viel Prozent Ihres Hausrates durch das Hochwasser zerstört wurde.</p> <p><u>zg6</u> - In welchem Jahr wurde das Gebäude erbaut?</p>
HW Typ2 gclass5	334	<p><u>ce1haupt</u> - Korrekturspalte neu: Fluss, der hauptsächlich für den Schaden</p> <p><u>cgs20oz</u> - Korrekturspalte: Teuerster Einzelschaden am Gebäude Wert in Euro (Zahl)</p> <p><u>chs11oz</u> - Korrekturspalte: Höhe der schon erhaltenen finanziellen Entschädigung</p> <p><u>e6aklasn</u> - Klassifizierung der Spalten e6a (Zusammenfassung aller Fälle wo 1. OG betroffen)</p> <p><u>erf7</u> - Seit wann wohnen Sie\ Ihre Familie schon in der vom Augusthochwasser 2002 betroffenen Wohnung?</p> <p><u>r2d</u> - Zeit für Reinigungsarbeiten in Tagen aus r2b und r2a/24h</p> <p><u>r3</u> - Wie viele Personen - sie selbst eingeschlossen - waren insgesamt mit den Reinigungsarbeiten am gesamten Gebäude beschäftigt? Nur das Saubermachen, NICHT die Renovierung.</p> <p><u>wstuegok</u> - Berechnung: Wasserstand über GOK (Kellerwerte minus 2.5 m)</p> <p><u>wstueke</u> - Berechnung: Wasserstand über Kellerboden</p> <p><u>zg6</u> - In welchem Jahr wurde das Gebäude erbaut?</p> <p><u>zgs16</u> - Wurden diese Arbeiten von Ihnen selbst durchgeführt, oder haben sie Handwerker beauftragt? Schätzen sie bitte, wie viel Prozent der angefallenen Gesamt-Arbeitszeit von Handwerkern erbracht wurde.</p>
HW Typ2 hclass5	477	<p><u>chs11oz</u> - Korrekturspalte: Höhe der schon erhaltenen finanziellen Entschädigung</p> <p><u>chs7oz</u> - Korrekturspalte: Wiederbeschaffungswert teuerstes Objekt in Euro</p> <p><u>e6aklasn</u> - Klassifizierung der Spalten e6a (Zusammenfassung aller Fälle wo 1. OG betroffen)</p> <p><u>erf7</u> - Seit wann wohnen Sie\ Ihre Familie schon in der vom Augusthochwasser 2002 betroffenen Wohnung?</p> <p><u>hs3az</u> - Wert des gesamten Hausrats in Euro</p> <p><u>hs6o</u> - Welches geschädigte Objekt von Ihrem Hausrat war Ihrer Meinung nach das Teuerste</p> <p><u>hs8</u> - Schätzen Sie bitte, wie viel Prozent Ihres Hausrates durch das Hochwasser zerstört wurde.</p> <p><u>vsuh_600</u> - VSU_Hausrat berechnet: CHS1 (Wohnfl.)*600 €</p>
HW Typ3 gclass5	385	<p><u>ce1haupt</u> - Korrekturspalte neu: Fluss, der hauptsächlich für den Schaden</p> <p><u>cgs20oz</u> - Korrekturspalte: Teuerster Einzelschaden am Gebäude Wert in Euro (Zahl)</p> <p><u>chs11oz</u> - Korrekturspalte: Höhe der schon erhaltenen finanziellen Entschädigung</p> <p><u>r2d</u> - Zeit für Reinigungsarbeiten in Tagen aus r2b und r2a/24h</p> <p><u>r3</u> - Wie viele Personen - sie selbst eingeschlossen - waren insgesamt mit den</p>

		<p>Reinigungsarbeiten am gesamten Gebäude beschäftigt? Nur das Saubermachen, NICHT die Renovierung.</p> <p><u>ss_h2a</u> - Schadensatz Hausrat (chs7aoz / cvsu_h2)</p> <p><u>wstuegok</u> - Berechnung: Wasserstand über GOK (Kellerwerte minus 2.5 m)</p> <p><u>zg6</u> - In welchem Jahr wurde das Gebäude erbaut?</p> <p><u>zgs16</u> - Wurden diese Arbeiten von Ihnen selbst durchgeführt, oder haben sie Handwerker beauftragt? Schätzen sie bitte, wie viel Prozent der angefallenen Gesamt-Arbeitszeit von Handwerkern erbracht wurde.</p>
HW Typ3 hclass5	417	<p><u>ce1haupt</u> - Korrekturspalte neu: Fluss, der hauptsächlich für den Schaden</p> <p><u>chs11oz</u> - Korrekturspalte: Höhe der schon erhaltenen finanziellen Entschädigung</p> <p><u>chs7oz</u> - Korrekturspalte: Wiederbeschaffungswert teuerstes Objekt in Euro</p> <p><u>czgs18oz</u> - Gebäudeschaden in Euro (ohne den höchsten Wert 4 Mio)</p> <p><u>e6bklasn</u> - Klassifizierung der Spalten e6b (Zusammenfassung aller Falle wo 1. OG betroffen)</p> <p><u>hs8</u> - Schätzen Sie bitte, wie viel Prozent Ihres Hausrates durch das Hochwasser zerstört wurde.</p> <p><u>zg6</u> - In welchem Jahr wurde das Gebäude erbaut?</p>
HW Typ4 gclass5	200	<p><u>ce1haupt</u> - Korrekturspalte neu: Fluss, der hauptsächlich für den Schaden</p> <p><u>cgs20oz</u> - Korrekturspalte: Teuerster Einzelschaden am Gebäude Wert in Euro (Zahl)</p> <p><u>erf7</u> - Seit wann wohnen Sie\ Ihre Familie schon in der vom Augusthochwasser 2002 betroffenen Wohnung?</p> <p><u>ev1</u> - Mussten Sie aufgrund des Hochwassers Ihre Wohnung bzw. Ihr Haus verlassen?</p> <p><u>gs1z</u> - Zusammenführung der Mieter und Eigentümerangaben, liegen beide Antworten vor, gilt die des Eigentümers! (keine Angabe/weiß nicht gelöscht)</p> <p><u>hs12</u> - Wie zufrieden waren Sie mit der Schadensregulierung?</p> <p><u>hs6o</u> - Welches geschädigte Objekt von Ihrem Hausrat war Ihrer Meinung nach das Teuerste</p> <p><u>kontjane</u> - Kontamination: ja/nein.</p> <p><u>ses_trad</u> - SES – Traditionell: Bildung, Beruf und Einkommen</p> <p><u>ss_h2a</u> - Schadensatz Hausrat (chs7aoz / cvsu_h2)</p> <p><u>v5</u> - Vorsorgemaßnahmen: Ich vermeide wertvolle fest installierte Inneneinrichtung in den hochwassergefährdeten Stockwerken, sondern verwende wasserwiderstandsfähige</p> <p><u>v9janein</u> - Vorsorgemaßnahmen: Ich schaffe stationäre oder mobile Wassersperren an, die verhindern, dass Wasser in das Gebäude\ die Wohnung eindringt (wie z.B. Schotten für Fenster</p> <p><u>wstuegok</u> - Berechnung: Wasserstand über GOK (Kellerwerte minus 2.5 m)</p> <p><u>zg5n</u> - Bauausführung, Spalten zg5 und g5o kombiniert</p> <p><u>zg6</u> - In welchem Jahr wurde das Gebäude erbaut?</p> <p><u>zgs21</u> - Wiederherstellung: Zusammenführung gs21, xgs21</p>
HW Typ4 hclass5	226	<p><u>chs7oz</u> - Korrekturspalte: Wiederbeschaffungswert teuerstes Objekt in Euro</p> <p><u>cs3</u> - Wohnen Sie zur Miete oder gehört Ihnen das Haus oder die Wohnung?</p> <p><u>e6bklasn</u> - Klassifizierung der Spalten e6b (Zusammenfassung aller Falle wo 1. OG betroffen)</p> <p><u>erf7</u> - Seit wann wohnen Sie\ Ihre Familie schon in der vom Augusthochwasser 2002 betroffenen Wohnung?</p> <p><u>hs3az</u> - Wert des gesamten Hausrats in Euro</p> <p><u>hs8</u> - Schätzen Sie bitte, wie viel Prozent Ihres Hausrates durch das Hochwasser zerstört wurde.</p>

		<u>mue_t</u> - Mündungsbereich, betroffene Gewässer
HW TypAll gclass5	107 9	<u>ce1haupt</u> - Korrekturspalte neu: Fluss, der hauptsächlich für den Schaden <u>cgs20oz</u> - Korrekturspalte: Teuerster Einzelschaden am Gebäude Wert in Euro (Zahl) <u>chs11oz</u> - Korrekturspalte: Höhe der schon erhaltenen finanziellen Entschädigung <u>e6aklasn</u> - Klassifizierung der Spalten e6a (Zusammenfassung aller Fälle wo 1. OG betroffen) <u>ev1</u> - Mussten Sie aufgrund des Hochwassers Ihre Wohnung bzw. Ihr Haus verlassen? <u>gs1z</u> - Zusammenführung der Mieter und Eigentümerangaben, liegen beide Antworten vor, gilt die des Eigentümers! (keine Angabe/weiß nicht gelöscht) <u>kontjane</u> - Kontamination: ja/nein. <u>r2d</u> - Zeit für Reinigungsarbeiten in Tagen aus r2b und r2a/24h <u>r3</u> - Wie viele Personen - sie selbst eingeschlossen - waren insgesamt mit den Reinigungsarbeiten am gesamten Gebäude beschäftigt? Nur das Saubermachen, NICHT die Renovierung. <u>ss_h2a</u> - Schadensatz Hausrat (chs7aoz / cvsu_h2) <u>wstgkl3</u> - Klassifizierung des Wasserstandes über GOK nach LTV Klassen <u>wstuegok</u> - Berechnung: Wasserstand über GOK (Kellerwerte minus 2.5 m) <u>wstueke</u> - Berechnung: Wasserstand über Kellerboden <u>zg6</u> - In welchem Jahr wurde das Gebäude erbaut? <u>zgs16</u> - Wurden diese Arbeiten von Ihnen selbst durchgeführt, oder haben sie Handwerker beauftragt? Schätzen sie bitte, wie viel Prozent der angefallenen Gesamt-Arbeitszeit von Handwerkern erbracht wurde.
HW TypAll hclass5	127 2	<u>chs11oz</u> - Korrekturspalte: Höhe der schon erhaltenen finanziellen Entschädigung <u>chs7oz</u> - Korrekturspalte: Wiederbeschaffungswert teuerstes Objekt in Euro <u>e6bklasn</u> - Klassifizierung der Spalten e6b (Zusammenfassung aller Fälle wo 1. OG betroffen) <u>hs10</u> - Bitte vergleichen Sie Ihren Hausrat VOR dem Hochwasser und seinen JETZIGEN Zustand miteinander. Haben Sie den Hausrat mittlerweile wieder ersetzt oder weist der Hausrat immer noch eklatante Mängel und Lücken infolge der Hochwassereinwirkung auf? <u>hs3az</u> - Wert des gesamten Hausrats in Euro <u>hs3bon</u> - Keine Angabe im SPSS-Variablenlabel <u>hs3w</u> - Wie groß ist die vom Hochwasser betroffene Wohnfläche? <u>hs8</u> - Schätzen Sie bitte, wie viel Prozent Ihres Hausrates durch das Hochwasser zerstört wurde. <u>ind_fw2</u> - Indikator: Bewertung Warninformation <u>r2d</u> - Zeit für Reinigungsarbeiten in Tagen aus r2b und r2a/24h <u>vsuh_600</u> - VSU_Hausrat berechnet: CHS1 (Wohnfl.)*600 € <u>wstuegok</u> - Berechnung: Wasserstand über GOK (Kellerwerte minus 2.5 m) <u>zg8</u> - Wie groß ist die Grundstücksfläche?

Tabelle 4.5 Ergebnisse der Feature Selection je Kombination aus Hochwassertyp und Schadenstyp

Die Anzahl der selektierten Attribute variiert zwischen 7 (bspw. HW Typ1 hclass5) und 16 (HW Typ4 gclass5). Dieses Ergebnis kommt zu Stande, weil von den 459 möglichen Attributen für HWTyp1, 543 für HWTyp2, 526 für HWTyp3 und 466 für HWTyp4 (siehe auch Tabelle 3.1) rund 450 nominale Attribute lediglich eine Liste der Listengröße 1 aufweisen. Die sehr hohe Anzahl der

nominalen Attribute, die nur einen einzigen Wert annehmen können, ist darauf begründet, dass bei der Betrachtung aller Daten unter den insgesamt 601 Attributen bereits 371 nominale Attribute mit der Listengröße 1 existieren (siehe auch Kapitel 3.2 - Datenbasis). Nach der Zerlegung des gesamten Datensatzes in vier Datensätze (je Hochwassertyp einer) erhöhte sich Anzahl der nominalen Attribute mit der Listengröße 1. All diese Attribute besitzen mit einer sehr hohen Wahrscheinlichkeit keine gute Aussagekraft in Hinblick auf das Zielattribut. Demzufolge bleiben pro Hochwassertyp insgesamt rund 10 bis 90 Attribute mit potentieller Aussagekraft übrig. Eine Anzahl von 7 bis 16 wirklich relevanten Attributen für die Vorhersage ist daher plausibel.

4.1.5 Gibt es Unterschiede zwischen den Hochwassertypen oder zwischen den Schadenstypen bzgl. der Güte der Vorhersage?

Am Beispiel für den Hochwassertyp 1 in der Abbildung 4-1 wird deutlich, dass es Unterschiede in der Güte der Vorhersage zwischen den beiden Schadenstypen gibt. Zwar ist für beide Schadenstypen die Vorhersage mit der Toleranz von einer Klasse Abweichung von der tatsächlichen Klasse ungefähr gleich gut (bei 90%), aber für die richtig vorhergesagte Klasse allein ist der Unterschied erheblich. So liegt für den Gebäudeschaden die relative Häufigkeit der korrekten Klassifikation bei 70% hingegen für den Hausratschaden bei nur 57% mit dem jeweils besten Algorithmus. Bei den Hochwassertypen 2 und 3 bestehen kaum Unterschiede in der Güte der Klassifikationsergebnisse je für beide Schadenstypen und beim Hochwassertyp 4 sind bessere Ergebnisse für den Gebäudeschaden zu verzeichnen.

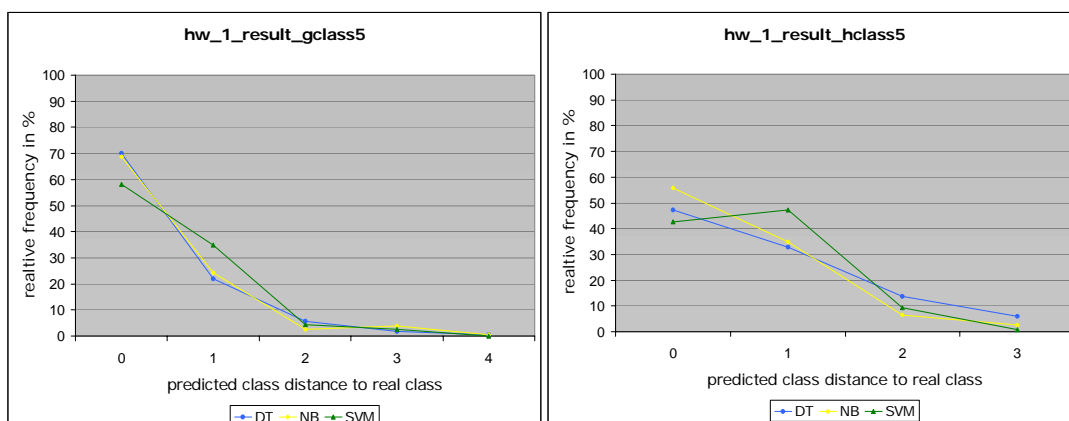


Abbildung 4-1 Klassifikationsergebnisse für den Hochwassertyp 1 und beide Schadenstypen. Der Vergleich zeigt ein wesentlich besseres Vorhersageergebnis für den Gebäudeschaden als für den Hausratschaden bezüglich der Menge der richtig klassifizierten Instanzen. Die Anzahl der Instanzen für die korrekt

vorhergesagte Klasse zusammen mit der Anzahl der Nachbarklasse als vorhergesagte Klasse ist bei beiden Schadenstypen ungefähr gleich. Entsprechend ist für die Toleranz der Abweichung um eine Klasse bei der Vorhersage, das Vorhersageergebnis nahezu gleich gut für beide Schadenstypen.

Der Vergleich der Hochwassertypen 3 und 4 für den Gebäudeschaden in der Abbildung 4-2 zeigt auf, dass auch zwischen den einzelnen Hochwassertypen für einen Schadenstyp Unterschiede in der Ergebnislösung bestehen. Während die relative Häufigkeit der richtig klassifizierten Instanzen beim Gebäudeschaden für den Hochwassertyp 4 bei 68% liegt, sind es beim Hochwassertyp 3 nur 51% mit dem besten Algorithmus.

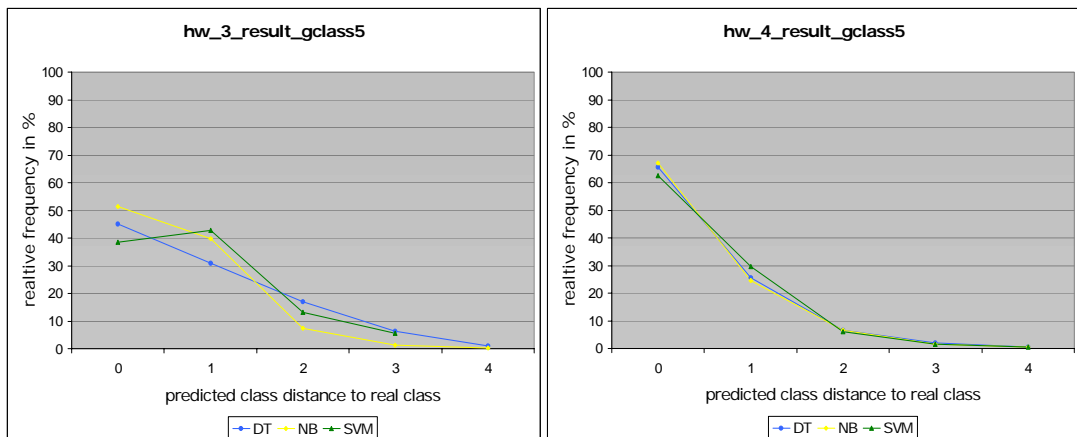


Abbildung 4-2 Klassifikationsergebnisse für den Hochwassertyp 3 und 4, je für den Gebäudeschaden. Der Vergleich innerhalb eines Schadenstyps zeigt, dass die Vorhersage abhängig vom Hochwassertyp unterschiedlich gut funktioniert. Hier ist das Vorhersageergebnis in Hinblick auf die richtig klassifizierten Instanzen für den Hochwassertyp 4 besser als für den Hochwassertyp 3.

Den übrigen Diagrammen im Anhang A-1 ist zu entnehmen, dass die Klassifikation des Schadens bei den Hochwassertypen 1 und 4 für den Gebäudeschaden und die Klassifikation des Schadens beim Hochwassertyp 4 für den Hausratschaden gute Vorhersageergebnisse (zwischen 60 und 72%) für die richtig klassifizierten Instanzen erzielt. Die Ergebnisse beim Hochwassertyp 2 sind innerhalb beider Schadenstypen im Vergleich zu den übrigen Hochwassertypen am schlechtesten ausgefallen.

Anmerkung

Bei der Ergebnisdarstellung in den Diagrammen handelt es sich im eigentlichen Sinne nicht um Kurven, da es keine Werte zwischen den einzelnen Abstandsklassen 0, 1, 2, 3 und 4 gibt und auch nicht geben kann. Dennoch wurde diese Darstellungsform gewählt, da sich die Unterschiede auf diese Art schneller visuell erfassen lassen, als wären nur Balkendiagramme verwendet worden.

4.1.6 Wird der Schaden generell eher unter- oder überschätzt?

Die „classified-as“ Matrizen im Anhang A-2 bilden die Grundlage, um Aussagen darüber treffen zu können, ob bei der Schadensvorhersage mit den verwendeten drei Klassifikationsmethoden der Schaden eher unter- oder überschätzt wurde. In diesen Matrizen ist die Information festgehalten, welche Schadensklasse für die Instanzen vorhergesagt wurde und welche Schadensklasse die richtige Klasse gewesen wäre. Dabei sind nur die Anzahlen der Instanzen je Klasse verzeichnet. Es gibt keine Rückschlussmöglichkeit auf die jeweils konkret betroffene Instanz. In der Tabelle 4.6 ist für die Klassifikation auf allen Daten je Schadenstyp die Unter- bzw. Überschätzung für alle Klassen zusammengefasst und auf die Abstände zu den richtigen Klassen reduziert. So entspricht die Anzahl der korrekt klassifizierten Instanzen den Angaben in den grün markierten Feldern. Links davon befinden sich die Angaben zur Anzahl der um eine Klasse unterschätzten Instanzen. Schrittweise weiter nach links gehend, ist die Anzahl der Instanzen notiert, deren vorhergesagte Klasse den Abstand um die Schrittlänge weiter weg zur richtigen Klasse hat. Die rechte Seite von den grünen Feldern ist für die überschätzten Instanzen analog zu interpretieren.

hwTypAllgclass5										hwTypAllhclass5									
DT										DT									
2	21	59	210	564	138	56	29	0		3	18	57	228	663	221	54	24	4	
292					223					306					303				
NB										NB									
2	7	37	237	608	151	24	13	0		4	28	91	356	611	135	39	6	2	
283					188					479					182				
SVM										SVM									
1	12	54	214	540	193	52	13	0		3	31	77	263	555	252	74	16	1	
281					258					374					343				

Tabelle 4.6 Anzahl unter- und überschätzter Instanzen für alle Daten je Schadenstyp und Klassifikationsmethode

Aus der Tabelle 4.6 geht hervor, dass bei beiden Schadenstypen allgemein mehr unterschätzt als überschätzt wurde. Bei den Klassifikationsmethoden DT und SVM besteht ein relativ ausgewogenes Verhältnis von der Anzahl der insgesamt unterschätzten zu den insgesamt überschätzten Instanzen:

- Gebäudeschaden DT: 292 zu 223, SVM: 281 zu 258 und
- Hausratschaden DT: 306 zu 303, SVM: 374 zu 343.

Hingegen dazu gibt es beim NB-Verfahren eine wesentlich größere Tendenz zum Unterschätzen (Gebäudeschaden: 283 zu 188, Hausratschaden: 479 zu 182). Aus den „classified-as“-

Matrizen im Anhang A-2 geht hervor, dass bei der Betrachtung der einzelnen Hochwassertypen ebenfalls generell eher unterschätzt wird. In lediglich 8 von 24 Matrizen ist die Gesamtanzahl der überschätzten Instanzen größer als die der unterschätzten Instanzen. Darunter viermal bei der Klassifikation mit dem NB und viermal bei der Klassifikation mit der SVM (je bei den Kombinationen aus Hochwassertyp und Schadenstyp: hwTyp2gclass5, hwTyp2hclass5, hwTyp4gclass5 und hwTyp4hclass5).

4.1.7 Wie verändern sich die Ergebnisse bei einer separaten Untersuchung von nur nominalen bzw. nur numerischen Attributen?

Die Algorithmen DT und SVM unterscheiden in ihren Berechnungen zwischen nominalen und numerischen Attributen, während der NB numerische Attribute in einem Vorbereitungsschritt in nominale konvertiert und folglich für seine Berechnungen nicht zwischen den beiden Datentypen unterscheidet. Es soll getestet werden, wie sich die unterschiedlichen Algorithmeigenschaften auf die Vorhersageergebnisse auswirken, indem die nominalen und numerischen Attribute getrennt voneinander betrachtet werden. Abbildung 4-3 repräsentiert dazu das Ergebnis einer beispielhaften Untersuchung der Klassifikation mit den Methoden DT, NB und SVM auf den Hochwasserdaten beim Hochwassertyp 1 für den Gebäudeschaden im Vergleich zu den entsprechenden originalen Ergebnissen. Original steht hierbei für das gemittelte Ergebnis über die drei angewendeten Algorithmen beim Hochwassertyp 1 für den Gebäudeschaden (hwTyp1gclass5) mit allen Attributen. Für die Unterscheidung nominal versus numerisch wurden die acht Attribute aus der Datei hwTyp1gclass5.arff in die Dateien hwTyp1_nominal_gclass5.arff (mit dem Anteil der nominalen Attribute) und hwTyp1_numeric_gclass5.arff (mit dem Anteil der numerischen Attribute) aufgesplittet. Die Aufteilung sieht demnach wie folgt aus:

hwTyp1_nominal_gclass5	hwTyp1_numeric_gclass5
<u>cs3</u> - Wohnen Sie zur Miete oder gehört Ihnen das Haus oder die Wohnung?	<u>cgs20oz</u> - Korrekturspalte: Teuerster Einzelschaden am Gebäude Wert in Euro (Zahl)
<u>erf7</u> - Seit wann wohnen Sie\ Ihre Familie schon in der vom Augusthochwasser 2002 betroffenen Wohnung?	<u>chs11oz</u> - Korrekturspalte: Höhe der schon erhaltenen finanziellen Entschädigung
<u>gs1z</u> - Zusammenführung der Mieter und Eigentümerangaben, liegen beide Antworten vor, gilt die des Eigentümers! (keine Angabe/weiß nicht gelöscht)	<u>wstuegok</u> - Berechnung: Wasserstand über GOK (Kellerwerte minus 2.5 m)
<u>zg6</u> - In welchem Jahr wurde das Gebäude erbaut?	<u>zgs16</u> - Wurden diese Arbeiten von Ihnen selbst durchgeführt, oder haben sie Handwerker beauftragt? Schätzen sie bitte, wie viel Prozent der angefallenen Gesamt-Arbeitszeit von Handwerkern erbracht wurde.

Zusätzlich existiert in beiden Dateien das Zielattribut gclass5.

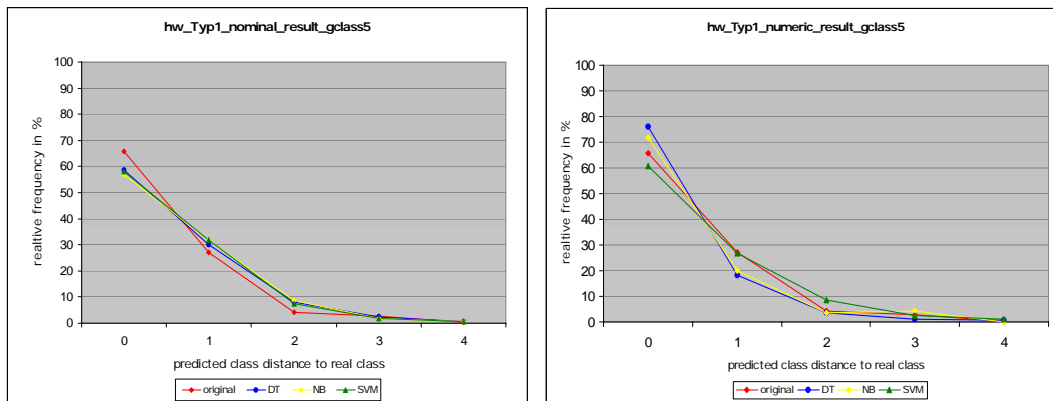


Abbildung 4-3 Vergleich der Klassifikationsergebnisse für den Hochwassertyp 1 nominal versus numerisch. Die Vorhersage auf den nominalen Attributen ist bei allen Algorithmen gleich gut, aber schlechter als auf der gesamten Attributmengende. Die Vorhersage auf den numerischen Attributen ist je Algorithmus unterschiedlich gut und gemittelt über alle Algorithmen gleich gut wie die Vorhersage auf der gesamten Attributmengende.

Das Klassifizieren auf den getrennten Attributmengen erzielte Vorhersageergebnisse, die Unterschiede in der Güte der Vorhersage zwischen den Algorithmen im Umgang mit numerischen Attributen aufzeigt. Die Ergebnisse auf der nominalen Attributmengende sind jedoch für alle Algorithmen gleich. Zwischen den Attributmengen gibt es in diesem Beispiel gemittelt über alle Algorithmen leicht bessere Ergebnisse für die Menge der numerischen Attribute.

Für den Datensatz der vorliegenden Hochwasserdaten bedeutet dies, dass die Anzahl der Attribute aus den Daten, die durch die Feature Selection gefiltert wurden und die gut auf die Zielklasse mappen

- entweder relativ gleich verteilt ist zwischen den nominalen und numerischen Attributen
- oder es weniger gut mappende Attribute bei den nominalen gibt als in numerischen vorhanden, dabei aber mappen diese weniger genauso gut auf die Zielklasse wie die anzahlmäßig überlegenen numerischen Attribute im Durchschnitt
- oder umgekehrt

und das mit einer leichten Tendenz zu mehr gut mappenden numerischen Attributen.

Im Allgemeinen besagt das Ergebnis, dass die Trennung der beiden Attributtypen zu leicht unterschiedlich guten Vorhersagen führen kann, als eine Mischung aus beiden. Dennoch gibt es auf keiner Seite der beiden Attributtypen ein Attribut oder mehrere Attribute die nennenswert besser auf die Zielklasse mappen als die Attribute vom jeweils anderen Attributtyp.

5 Zusammenfassung und Ausblick

Vom GeoForschungsZentrum in Potsdam wurden für diese Arbeit 1697 Datensätze zum Flutereignis an Elbe und Donau im August 2002 bereitgestellt. Diese Daten sind mit ihren über 600 Attributen zu komplex, um alle flutbegleitenden Faktoren visuell zu überblicken und dabei Aussagen über ihren Einfluss auf die entstandenen Schäden an den entsprechenden Ein- bzw. Mehrfamilienhäusern treffen zu können. Deshalb erfolgte in dieser Arbeit erstmalig eine Analyse der Daten mittels Methoden des maschinellen Lernens. Prototypisch wurde ein Prozess entwickelt, mit dem drei verschiedene Klassifikationsalgorithmen getestet und deren Vorhersageergebnisse evaluiert wurden.

Das Vorgehen der Analyse beginnt mit dem Filtern der besonders diskriminativen Eigenschaften für die entstandenen Gebäude- bzw. Hausratschäden unter Berücksichtigung der vier verschiedenen Hochwassertypen „Grundhochwasser“, „langsame Flussüberschwemmung“, „Sturzflut“ und „Deichbruch“. Die ermittelte Selektion der Attribute hat sich größtenteils als abhängig von den jeweiligen Hochwassertypen aber auch als abhängig vom jeweiligen Schadenstyp herausgestellt. Nur sehr wenige Attribute tauchen in allen berechneten Feature Selections auf.

Basierend auf den gefilterten Eigenschaften werden nacheinander die Klassifikationsmethoden angewendet. Das Trainieren der Klassifikationsmodelle erfolgte mit dem C4.5 als Algorithmus des Entscheidungsbaumlernens, mit dem Naiv Bayes Klassifikator und mit einer Support Vector Machine. Sowohl der Vorgang der Feature Selection als auch das Klassifizieren an sich sind innerhalb einer Software für maschinelles Lernen und Data Mining namens Weka durchgeführt worden. Dieses Tool erfordert ein spezielles Eingabeformat (ARFF), so dass die originalen Daten aus der Software SPSS stammend in dieses Eingabeformat konvertiert werden mussten. Eigens entwickelte SPSS- und Perlskripte übernehmen die Funktion der Konvertierung der Eingangsdaten und die Evaluierung der Klassifikationsergebnisse, die von dem Tool Weka generiert wurden.

Ein Vergleich der vorhergesagten Schadensklassen mit den tatsächlichen Schadensklassen ergab, dass im Durchschnitt über alle Methoden und deren Ergebnisse die richtige Klasse zu 50% und die richtige Klasse oder deren Nachbarklasse zu 86% vorhergesagt wurde. Welcher Algorithmus jeweils beste Ergebnisse erzielte, hängt auch von der Art des Schadens und dem Hochwassertyp ab. Bei der Klassifikation der Gebäudeschäden hat sich der Naiv Bayes Klassifikator als beste Methode herauskristallisiert. Für die Hausratschäden erzielte der Naiv Bayes Klassifikator für die Hochwassertypen „Grundhochwasser“ und „langsame Flussüberschwemmung“ die besten Ergebnisse. Das Entscheidungsbaumlernen mit dem C4.5 zeigte sich als

die beste Methode für die Schadensvorhersage bei den Hochwassertypen „Sturzflut“ und „Deichbruch“. Generell hat die Vorhersage für den Gebäudeschaden bessere Ergebnisse erzielt als für den Hausratschaden. Für die Hochwassertypen „Grundhochwasser“ und „langsame Flussüberschwemmung“ ergab die Vorhersage bessere Ergebnisse als für die Hochwassertypen „Sturzflut“ und „Deichbruch“.

Für das Trainieren der Klassifikationsmodelle standen zwar 1697 Datensätze zur Verfügung, davon konnten jedoch nur 1079 Datensätze für die Vorhersage des Gebäudeschadens und nur 1272 Datensätze für die Vorhersage des Hausratschadens verwendet werden, da für den Rest die Klasseninformation (der jeweilige Schadensbetrag) fehlte. Ausblickend wäre zu untersuchen ob mehr Datensätze die bisherigen Aussagen bestätigen würden oder sogar eine Ergebnisverbesserung zur Folge hätten. Weiterhin wäre zu prüfen ob basierend auf einer größeren Datenmenge eine feinere Klasseneinteilung in mehr als fünf Klassen weiterhin gute Vorhersagen erzielen würde. Genauer gilt es herauszufinden, wie eine optimale Klassenbreite aussehen würde, die dem Kompromiss zwischen bestmöglicher Vorhersage und kleinstmöglicher Klasseneinteilung gerecht wird. Eine zusätzliche Verbesserung der Klassifikationsergebnisse könnte im Justieren der vielen Parameter bestehen, die Weka für Feature Selections und zum Klassifizieren bereitstellt. Zudem könnten verschiedene Selektionsmethoden angewendet und hinsichtlich einer potentiellen Ergebnisoptimierung verglichen werden.

Literatur- und Quellenverzeichnis

[Eckstein, 2006] Silke Eckstein: Text-Data-Mining, verfügbar: http://infbsdb1.idb.cs.tu-bs.de/eckstein/ISBI0607/isbi0607_5_Text-Data-Mining.pdf, (August 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)

[Hall and Smith, 1997] Mark A. Hall and Lloyd A. Smith, A Correlation Based Filter Approach, *Department of Computer Science, University of Waikato*, 1997

[Hall, 1998] Mark A. Hall, Correlation-based Feature Subset Selection for Machine Learning, *Thesis submitted in partial fulfilment of the requirements of the degree of Doctor of Philosophy at the University of Waikato*, 1998

[Hall and Smith, 1998] Mark A. Hall and Lloyd A. Smith, Practical Feature Subset Selection for Machine Learning, *Department of Computer Science, University of Waikato*, 1998

[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

[John and Langley, 1995] Gorge H. John and Pat Langley: Estimating Continuous Distributions in Bayesian Classifiers, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers*, 1995

[Keerthi et. al, 2001] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, K. R. K. Murthy: Improvements to Platt's SMO Algorithm for SVM Classifier Design, *Neural Computation*, 13(3), 2001

[Last, 2008] Mark Last: Dimensionality Reduction and Feature Selection, verfügbar: <http://www.ise.bgu.ac.il/courses/dw/slides/dw9.pdf>, (August 2008)

[Liu and Motoda, 1998] Huan Liu and Hiroshi Motoda: Feature Selection for Knowledge Discovery and Data Mining, *Springer*, 1998

[Mitchell, 1997] Tom M. Mitchell: Machine Learning, *McGraw-Hill International Editions*, 1997

[Platt, 1998] J. Platt: Fast Training of Support Vector Machines using Sequential Minimal Optimization. *Advances in Kernel Methods – Support Vector Learning*, B.Schoelkopf, C. Burges, and A. Smola, eds., *MIT Press*, 1998

- [Quinlan, 1993] Ross Quinlan: Programs for Machine Learning, *Morgan Kaufmann Publishers*, 1993
- [Scheffer, 2004] Tobias Scheffer: Hypothesenbewertung, verfügbar: <http://www2.informatik.hu-berlin.de/wm/mlm2004/Hypothesenbewertung.pdf>, (August 2008)
- [Scheffer, 2006] Tobias Scheffer: Maschinelles Lernen und Data Mining, verfügbar: <http://www2.informatik.hu-berlin.de/wm/mlm2006w.html>, (August 2008)
- [Thielen et. al, 2005] A. H. Thielen, 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
- [Thielen et. al, 2007] A. H. Thielen, 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
- [Weka, 2008] Data Mining Software in Java, verfügbar: <http://www.cs.waikato.ac.nz/ml/weka/>, (August 2008)
- [Weka ARFF, 2008] Attribute-Relation File Format (ARFF), verfügbar: <http://www.cs.waikato.ac.nz/~ml/weka/arff.html>, (August 2008)
- [Yu and Liu, 2004] Lei Yu and Huan Liu, Efficient Feature Selection via Analysis of Relevance and Redundancy, *Journal of Machine Learning Research*, October 2004

Anhang

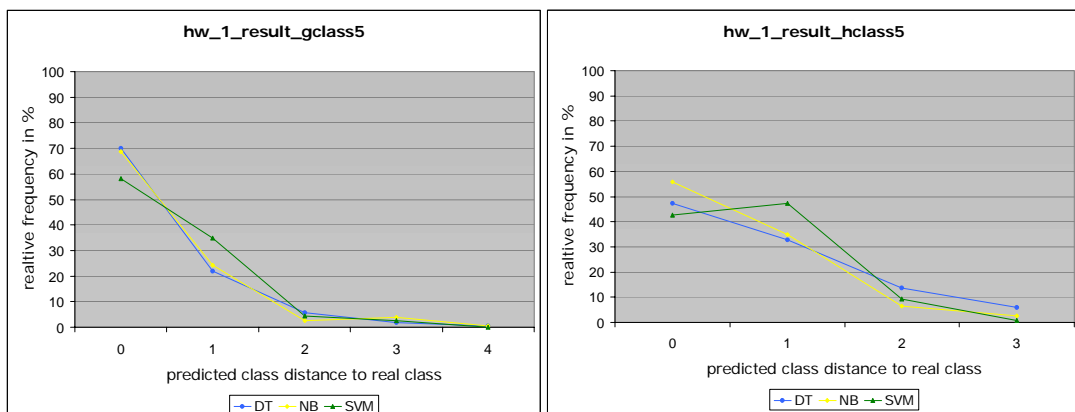
A Ergebnisauswertung

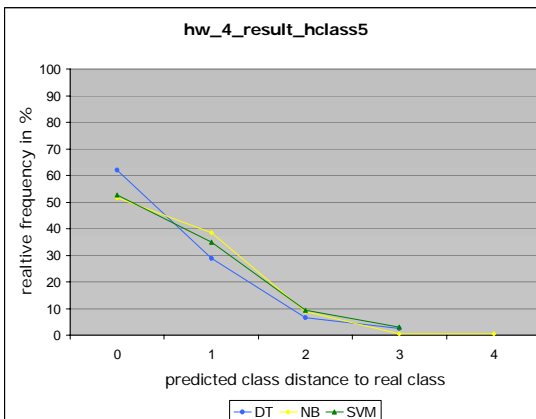
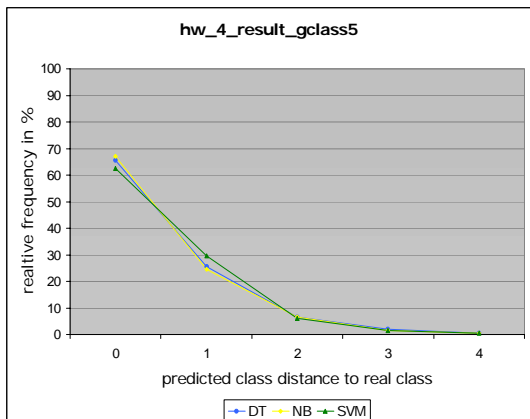
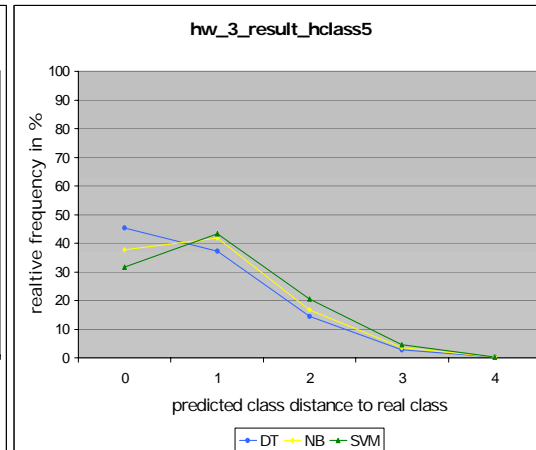
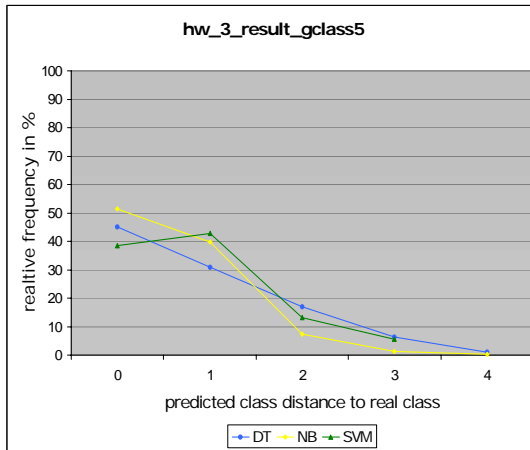
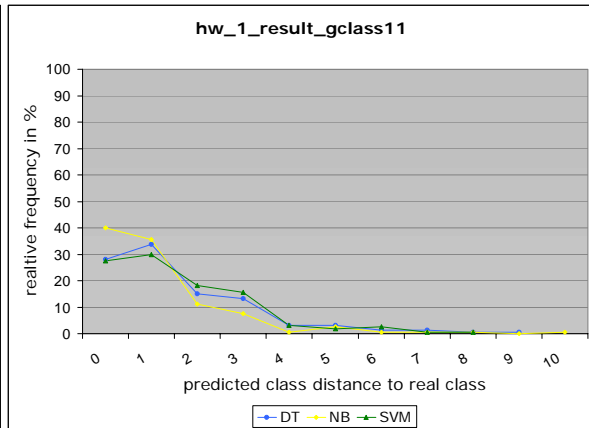
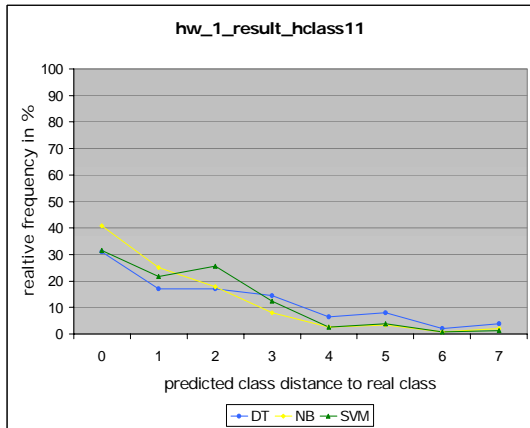
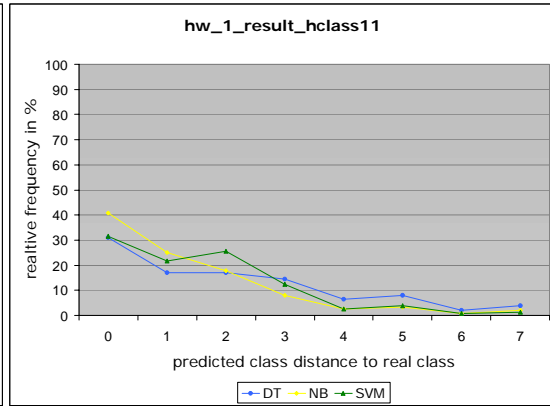
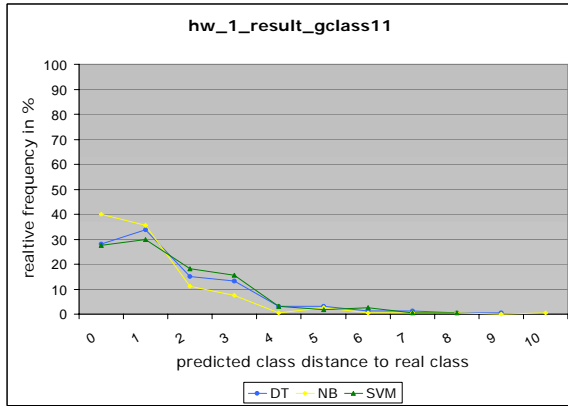
A-1 Klassifikationsergebnisdiagramme

Ein Diagramm enthält jeweils die Vorhersageergebnisse für die angewendeten Algorithmen aus den Bereichen des Entscheidungsbaum-Lernens (DT), des Bayes'sches Lernens (NB) und des Lernens mit einer Support Vector Machine (SVM). Die Vorhersageergebnisse werden in jedem Diagramm über die relative Anzahl der Instanzen pro Abstand der vorhergesagten Klasse von der echten Klasse repräsentiert.

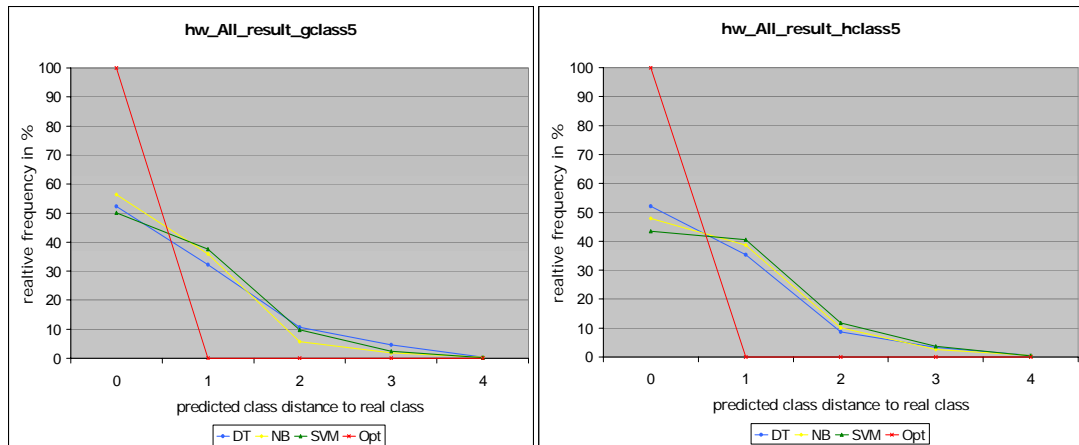
Die nebeneinander angeordneten Diagramme beinhalten die Ergebnisse für den Gebäudeschaden (links) und den Hausratschaden (rechts) für denselben Hochwassertyp.

Die untereinander angeordneten Diagramme beinhalten die Ergebnisse für die verschiedenen Hochwassertypen 1 bis 4 und „all“ (alle 4 Hochwassertypen zusammen).





Die Kurve „Opt“ steht für den theoretisch optimalen Kurvenverlauf.



A-2 Classified-as Matrizen

Die classified-as Matrizen geben die jeweilige Anzahl der Instanzen wieder, die als Klasse a, b, c, d oder e klassifiziert wurden (Spalten), wobei die echte Klasse die Klasse a, b, c, d oder e gemäß der entsprechenden Zeile ist. Je Kombination aus Hochwassertyp, Schadenstyp und Algorithmus existiert eine Matrix.

Die Matrizen sind wie folgt zu lesen:

- Die angegebene Anzahl der Instanzen in den Matrixzellen ist eine absolute Angabe.
- In der Diagonale einer Matrix stehen die jeweils richtig klassifizierten Instanzen.
- Von der Diagonale um x Schritte weiter nach links bzw. nach rechts ist entsprechend die Anzahl der Instanzen zu finden, deren vorhergesagte Klasse um x Klassen entfernt ist von der echten Klasse.
- All jene Matrixeinträge unterhalb der Diagonalen entsprechen einer unterschätzten Vorhersage.
- All jene Matrixeinträge oberhalb der Diagonalen entsprechen einer überschätzten Vorhersage.
- Hinter „a =“ bis „e =“ steht die jeweilige Klassenspanne des Schadens in von bis Euro-Beträgen.

HW Typ	Schadenstyp	Confusion Matrix																	
		DT					NB					SVM							
1	Gebäude	a	b	c	d	e	<-- classified as	a	b	c	d	e	<-- classified as	a	b	c	d	e	<-- classified as
		90	7	0	0	0	a = 0-5000	86	10	0	0	1	a = 0-5000	80	15	1	1	0	a = 0-5000
		18	20	0	0	0	b = 5001-15000	14	18	4	0	2	b = 5001-15000	20	12	5	1	0	b = 5001-15000
		3	10	2	0	0	c = 15001-30000	1	6	4	2	2	c = 15001-30000	2	11	1	1	0	c = 15001-30000
		0	6	1	0	0	d = 30001-100000	1	1	3	2	0	d = 30001-100000	2	2	3	0	0	d = 30001-100000
		1	2	0	0	0	e = 100001-755000	0	2	0	1	0	e = 100001-755000	0	0	1	2	0	e = 100001-755000
1	Hausrat	a	b	c	d	<-- classified as	a	b	c	d	<-- classified as	a	b	c	d	<-- classified as			
		72	0	0	0	a = 0-1900	54	14	3	1	a = 0-1900	45	25	2	0	a = 0-1900			
		50	0	0	0	b = 1901-5000	20	23	6	0	b = 1901-5000	26	19	4	0	b = 1901-5000			
		21	0	0	0	c = 5001-11000	5	9	6	2	c = 5001-11000	9	12	1	0	c = 5001-11000			
		9	0	0	0	d = 11001-25000	3	1	3	2	d = 11001-25000	1	4	4	0	d = 11001-25000			
2	Gebäude	a	b	c	d	e	<-- classified as	a	b	c	d	e	<-- classified as	a	b	c	d	e	<-- classified as
		0	0	0	40	0	a = 0-5000	33	2	2	3	0	a = 0-5000	13	14	7	6	0	a = 0-5000
		0	0	0	67	0	b = 5001-15000	28	23	10	3	3	b = 5001-15000	11	26	17	13	0	b = 5001-15000
		0	0	0	74	0	c = 15001-30000	12	18	21	22	1	c = 15001-30000	3	12	22	33	4	c = 15001-30000
		0	0	0	114	0	d = 30001-100000	5	7	18	69	15	d = 30001-100000	5	9	20	66	14	d = 30001-100000
		0	0	0	39	0	e = 100001-755000	0	1	2	17	19	e = 100001-755000	1	0	3	26	9	e = 100001-755000
2	Hausrat	a	b	c	d	e	<-- classified as	a	b	c	d	e	<-- classified as	a	b	c	d	e	<-- classified as
		19	60	2	5	2	a = 0-1900	66	19	1	0	2	a = 0-1900	28	39	6	5	10	a = 0-1900
		12	81	2	14	7	b = 1901-5000	47	56	6	5	2	b = 1901-5000	20	62	9	17	8	b = 1901-5000
		7	29	1	11	18	c = 5001-11000	16	23	10	13	4	c = 5001-11000	14	22	9	11	10	c = 5001-11000
		2	31	5	15	44	d = 11001-25000	6	20	11	49	11	d = 11001-25000	2	22	5	26	42	d = 11001-25000
		4	19	6	8	73	e = 25001-160000	3	12	7	39	49	e = 25001-160000	4	12	5	30	59	e = 25001-160000
3	Gebäude	a	b	c	d	e	<-- classified as	a	b	c	d	e	<-- classified as	a	b	c	d	e	<-- classified as
		43	12	4	29	0	a = 0-5000	33	25	26	4	0	a = 0-5000	34	19	19	13	3	a = 0-5000
		24	3	9	37	0	b = 5001-15000	8	30	27	8	0	b = 5001-15000	18	21	12	22	0	b = 5001-15000
		16	10	9	50	0	c = 15001-30000	3	13	19	45	5	c = 15001-30000	7	10	20	41	7	c = 15001-30000
		4	2	1	19	0	d = 30001-100000	1	1	0	19	5	d = 30001-100000	0	4	3	17	2	d = 30001-100000
							e = 100001-755000										e = 100001-755000		
3	Hausrat	a	b	c	d	e	<-- classified as	a	b	c	d	e	<-- classified as	a	b	c	d	e	<-- classified as
		40	29	0	4	0	a = 0-1900	59	11	2	1	0	a = 0-1900	7	51	8	6	1	a = 0-1900
		5	83	3	27	1	b = 1901-5000	62	38	11	7	1	b = 1901-5000	16	72	11	19	1	b = 1901-5000
		3	30	8	38	4	c = 5001-11000	24	23	24	10	2	c = 5001-11000	7	30	13	26	7	c = 5001-11000
		2	21	13	46	7	d = 11001-25000	8	27	23	15	16	d = 11001-25000	4	36	10	30	9	d = 11001-25000
		1	5	4	31	12	e = 25001-160000	1	5	9	17	21	e = 25001-160000	0	8	7	28	10	e = 25001-160000
4	Gebäude	a	b	c	d	e	<-- classified as	a	b	c	d	e	<-- classified as	a	b	c	d	e	<-- classified as
		0	0	0	4	0	a = 0-5000	2	0	1	1	0	a = 0-5000	2	0	1	1	0	a = 0-5000
		0	0	0	13	0	b = 5001-15000	1	2	4	5	1	b = 5001-15000	0	1	4	7	1	b = 5001-15000
		0	0	0	25	0	c = 15001-30000	0	2	10	13	0	c = 15001-30000	0	4	5	15	1	c = 15001-30000
		0	0	0	131	0	d = 30001-100000	0	4	8	113	6	d = 30001-100000	0	3	10	109	9	d = 30001-100000
		0	0	0	27	0	e = 100001-755000	0	1	3	16	7	e = 100001-755000	0	1	0	18	8	e = 100001-755000
4	Hausrat	a	b	c	d	e	<-- classified as	a	b	c	d	e	<-- classified as	a	b	c	d	e	<-- classified as
		0	3	0	1	0	a = 0-1900	3	0	1	0	0	a = 0-1900	0	2	1	1	0	a = 0-1900
		0	10	0	8	2	b = 1901-5000	6	3	2	9	0	b = 1901-5000	0	5	4	6	5	b = 1901-5000
		0	4	1	12	1	c = 5001-11000	0	5	3	6	4	c = 5001-11000	1	2	4	4	7	c = 5001-11000
		0	5	0	29	37	d = 11001-25000	1	3	2	50	15	d = 11001-25000	0	2	2	21	6	d = 11001-25000
		0	3	0	10	100	e = 25001-160000	1	0	3	50	59	e = 25001-160000	0	1	3	20	89	e = 25001-160000
All	Gebäude	a	b	c	d	e	<-- classified as	a	b	c	d	e	<-- classified as	a	b	c	d	e	<-- classified as
		197	23	5	29	0	a = 0-5000	216	24	5	9	0	a = 0-5000	176	47	20	11	0	a = 0-5000
		47	81	27	51	0	b = 5001-15000	69	65	52	16	4	b = 5001-15000	53	63	57	31	2	b = 5001-15000
		21	61	18	87	0	c = 15001-30000	14	65	57	48	3	c = 15001-30000	18	57	41	70	1	c = 15001-30000
		14	37	17	268	1	d = 30001-100000	5	21	46	238	27	d = 30001-100000	12	27	43	236	19	d = 30001-100000
		2	7	1	85	0	e = 100001-755000	2	2	2	57	32	e = 100001-755000	1	0	9	61	24	e = 100001-755000
All	Hausrat	a	b	c	d	e	<-- classified as	a	b	c	d	e	<-- classified as	a	b	c	d	e	<-- classified as
		176	43	3	11	4	a = 0-1900	200	23	10	2	2	a = 0-1900	127	96	7	6	1	a = 0-1900
		58	169	43	21	13	b = 1901-5000	139	109	31	21	4	b = 1901-5000	101	137	7	49	10	b = 1901-5000
		7	66	31	55	30	c = 5001-11000	34	47	61	39	8	c = 5001-11000	25	72	14	60	18	c = 5001-11000
		4	37	46	99	80	d = 11001-25000	16	37	50	121	42	d = 11001-25000	14	43	18	102	89	d = 11001-25000
		3	14	13	58	188	e = 25001-160000	4	12	20	120	120	e = 25001-160000	3	17	9	72	175	e = 25001-160000