KI-93: 17. Fachtagung für Künstliche Intelligenz Berlin, 13.-16. September 1993

Workshop

"Maschinelles Lernen: Theoretische Ansätze und Anwendungsaspekte"

GINI-ENTROPIE: Ein Vergleich zweier Heuristiken für den ID3/C4-Induktionsalgorithmus anhand verrauschter Grundwasserdaten

Harald Hiessl

Jürgen Lang

Fraunhofer-Gesellschaft
Institut für Systemtechnik und Innovationsforschung (FhG-ISI)
D-76139 Karlsruhe, Breslauer Str. 48, Tel. (0721) 8609-0

1. Fragestellung und Anwendungsbeispiel

Im Rahmen eines Forschungsprojektes wurde in einem Untersuchungsgebiet ein Grundwassergütemeßnetz eingerichtet (Grimm-Strele et al., 1991). An den Grundwasserbeobachtungsbrunnen des Meßnetzes wurde das Grundwasser beprobt und auf 22 verschiedene Beschaffenheitsparameter (stetige Attribute) untersucht. Mit Hilfe eines numerischen Strömungsmodells konnten aufgrund der Strömungsverhältnisse und der Grundwasserneubildung für 63 Beobachtungsbrunnen die Eintragsflächen bestimmt werden. Die Landnutzungsarten in diesen Eintragsflächen wurden von Fachleuten in 5 Nutzungsklassen eingeteilt. Folgende Nutzungsklassen wurden unterschieden:

Klasse 1: Belastungsarme Landnutzung
Klasse 2: Bebaute Gebiete, Mischgebiete
Klasse 3: Landwirtschaftliche genutzte Gebiete
Klasse 4: Durch Oberflächengewässer beeinflußte Gebiete
Klasse 5: Sonderbelastungen, Deponien

Es stellte sich die Frage, ob es eine (möglichst einfache und eindeutige) Beziehung zwischen dem Landnutzungstyp im Eintragsbereich eines Brunnens und der im Brunnen beobachteten Grundwasserbeschaffenheit gibt. Falls eine solche Beziehung gefunden werden kann, so kann sie beispielsweise dazu dienen,

- diejenigen Grundwasserbeschaffenheitsparameter zu identifizieren, die für die Eintragsflächen besonders typisch sind (z. B. zum Entwurf nutzungsspezifischer Meßprogramme zwecks Reduktion des Meßaufwandes),

- die Meßstelle anhand des beobachteten Grundwassertyps einer Grundwasserlandschaft zuzuordnen,

- mit Hilfe des Landnutzungstyps einer Eintragsfläche gewisse charakteristische Beschaffenheitsmuster des Grundwassers zu prognostizieren (z. B. zur Außreißerkontrolle oder zur Abschätzung der Auswirkungen geplanter Nutzungsänderungen auf die Grundwasserbeschaffenheit, oder bei nicht beprobten Brunnen mit bekanntem Nutzungstyp der zugehörigen Eintragsflächen die zu erwartende Grundwasserbeschaffenheit abzuschätzen) oder

- anhand von Messungen der Grundwasserbeschaffenheit auf die das Grundwasser maßgeblich beeinflussenden Nutzungen im bisher nicht typisierten Eintragsgebiet der Beobachtungsbrunnen zu schließen.

Im Rahmen des o.g. Forschungsprojektes wurde zum ersten Mal versucht, Beziehungen zwischen der Grundwasserbeschaffenheit und der typischen Nutzung in den Eintragsflächen der jeweiligen Beobachtungsbrunnen mit Hilfe von Induktionsmethoden abzuleiten.

2. Regelinduktion

2.1. Der ID3/C4-Algorithmus

Im vorliegenden Fall wird eine Re-Implementierung des C4-Algorithmus von Quinlan (1983, 1986) als Induktionsverfahren verwendet. Der C4-Algorithmus stellt eine Weiterentwicklung des "Iterative Dichotomizer 3 (ID3)"-Algorithmus (Quinlan, 1979) in Bezug auf das Fensterverfahren und die Bearbeitbarkeit kontinuierlicher Attribute dar. Beide Algorithmen verwenden den CLS ("Concept Learning System")-Algorithmus von Hunt als das eigentliche Induktionsverfahren.

Der CLS-Algorithmus leitet aus einer nicht zu großen Trainingsmenge von Beispielen rekursiv einen entsprechenden Regelsatz zur vollständigen Klassifikation der Objekte in Form eines Entscheidungsbaumes ab. Dabei durchläuft er im wesentlichen die folgenden drei Arbeitsschritte:

CLS-Schritt 1: Wenn alle Objekte einer Trainingsmenge T einer Klasse zugehören, dann erzeuge für diese Klasse einen Knoten. Stop. Andernfalls wähle (durch Anwendung eines geeigneten heuristischen Kriteriums) ein Attribut A_i (i aus $\{1,...,N\}$) aus und erzeuge für jeden von n Wertebereichen des Attributes einen entsprechenden Entscheidungsknoten.

CLS-Schritt 2: Teile die Trainingsmenge T entsprechend den Wertebereichen in n Teilmengen T_1 bis T_n.

CLS-Schritt 3: Wende den Algorithmus rekursiv auf jede der Teilmengen T_j (j aus $\{1,...,n\}$) an.

Quinlan verwendete als heuristisches Auswahlkriterium im CLS-Schritt 1 die Shannon'sche Informationsentropie. In der hier vorliegenden Arbeit wurde als alternative Heuristik noch das GINI-Kriterium (Breiman et al., 1984) implementiert. Eine kurze Darstellung der beiden implementierten Heuristiken ist im folgenden Abschnitt zu finden.

Der CLS erfordert, daß alle Objekte in der Trainingsmenge während des CLS-Schrittes 1 im wahlfreien Zugriff verfügbar sind. Dies setzt dem Umfang des Lernproblems, das mit CLS gelöst werden kann praktische Grenzen. In diesem Sinne stellt der ID3/C4-Algorithmus eine Erweiterung des CLS auf Lernprobleme mit umfangreichen Trainingsdatensätzen dar. Während der ursprüngliche ID3-Algorithmus nur diskrete Attribute verarbeiten konnte, wurde in der hier durchgeführten Implementierung auf die Verarbeitung von Attributen mit stetigem Wertebereich erweitert. Darüberhinaus wurde noch die Möglichkeit zum Pruning, d.h. Entfernung statistisch nicht signifikanter Zweige, des induzierten Entscheidungsbaumes vorgesehen, um das sog. "overfitting" des Entscheidungsbaumes zu vermeiden. Als Pruningverfahren wurde das von Quinlan (1987) vorgeschlagene "Pessimistic Pruning" gewählt.

Abb.1 zeigt die hier verwendete Implementation des ID3/C4-Algorithmus in Fließdiagrammform.

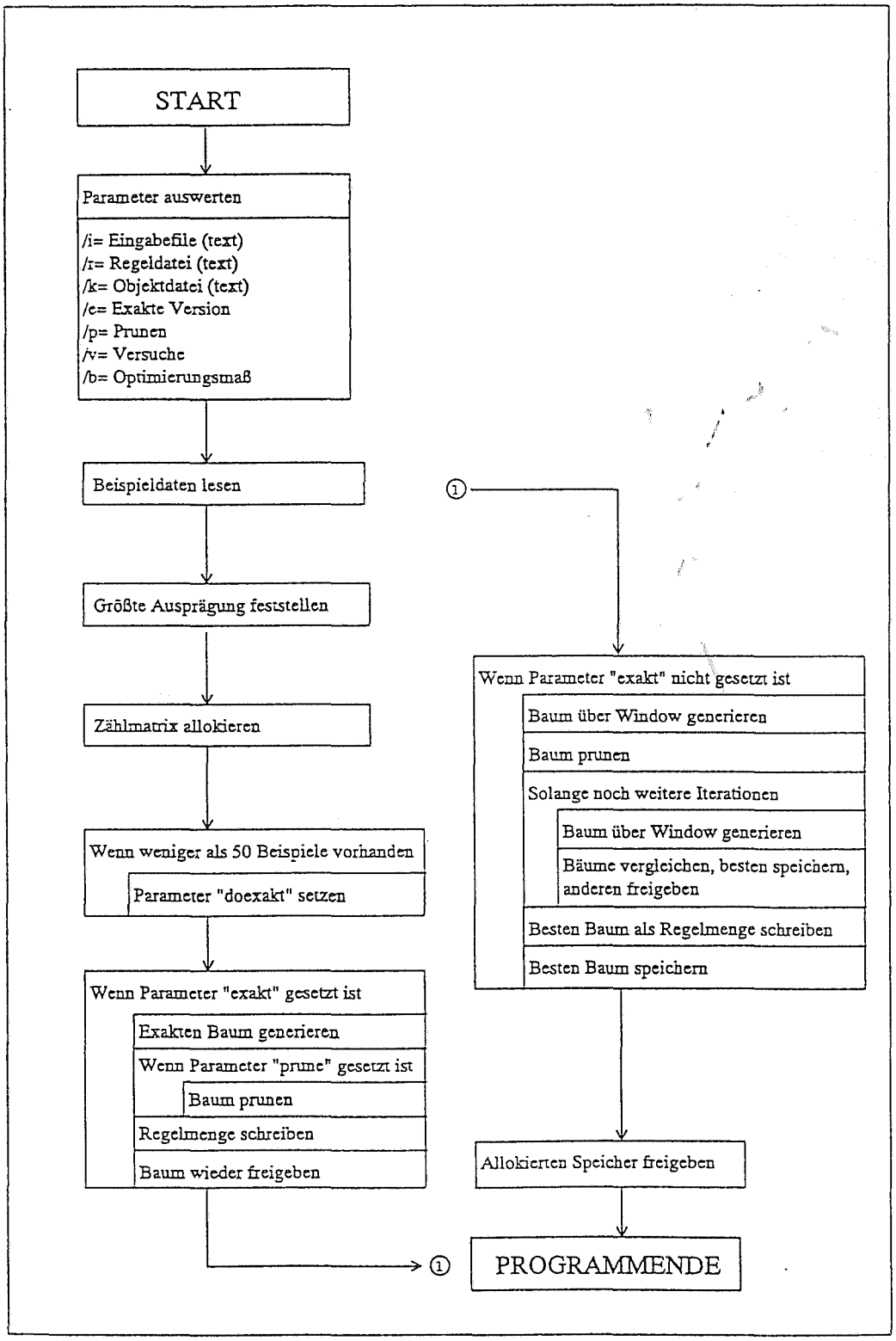

Abb.1: Fließdiagramm des ID3/C4-Algorithmus.

2.2. Entropie- und GINI-Kriterium: Zwei alternative Heuristiken für ID3/C4

Bei der Induktion des Klassifikationsbaumes sind sequentiell Entscheidungen zu treffen, welches Attribut bzw. welcher Attributwert im jeweiligen Berechnungsschritt für die weitere Konstruktion des Klassifikationsbaumes ausgewählt werden soll (Problem der Attributauswahl, Problem der Binarisierung stetiger Attribute). Grundsätzlich wird man bestrebt sein, Attribute bzw. Intervalle von Attributwerten auszuwählen, die im derzeitigen Berechnungsstadium des Baumes die noch nicht klassifizierten Objekte am besten klassifizieren, d.h. am besten zwischen den Klassen diskriminieren. Aus diesem Grunde müssen die im jeweiligen Berechnungsschritt des Baumes verfügbaren Attribute bzw. die möglichen Intervallgrenzen für stetige Attribute mit Hilfe eines Kriteriums hinsichtlich ihrer "Diskriminationsfähigkeit" bewertet werden.

In der vorliegenden Implementierung des ID3/C4-Algorithmus wurden zwei verschiedene Auswahlkriterien (Auswahlheuristiken) zur Verfügung gestellt. Zum einen wurde das in der Urfassung des ID3-Algorithmus von Quinlan verwendete Entropie-Kriterium implementiert (Quinlan, 1979) und zum anderen das von Breiman et al. (1984) zuerst für Machine-Learning Verfahren eingesetzte GINI-Kriterium.

In der Statistik werden derartige Maße als "Konzentrationsmaße" (Marfels, 1977) bezeichnet, wobei im Zusammenhang mit Machine-Learning unter "Konzentration" die Konzentration der für die Klassifikation relevanten Information auf bestimmte Attribute bzw. Intervalle von Attributwerten zu verstehen ist. Während das Entropiekriterium zur Gruppe der Maße für die absolute Konzentration gehört, ist das GINI-Kriterium den Maßen für die relative Konzentration zuzuordnen. Bei den absoluten Konzentrationsmaßen werden die Zahl der Merkmalsträger und ihre Größe in bestimmter Weise miteinander verknüpft und so die Ballung der Merkmalswerte auf absolut wenige Einheiten gemessen. Sie werden aus diesem Grunde als Konzentrationsmaße (i.e.S.) bezeichnet. Die relativen Konzentrationsmaße stellen dagegen die Ungleichheit in der Größe der Einheiten dar, wobei ihre Zahl belanglos ist. Sie werden aus diesem Grunde auch als Disparitätsmaße bezeichnet.

Ein empirischer Vergleich der beiden Heuristiken Entropie und GINI (neben weiteren Heuristiken) im Zusammenhang mit Machine-Learning Verfahren wurde von Mingers (1989) durchgeführt. Mingers stellte dabei fest, daß die Güte eines induzierten (nicht geprunten oder geprunten) Klassifikationsbaumes bei seiner Anwendung auf bisher nicht gesehene Objekte (prediction accuracy) nicht signifikant von der verwendeten Auswahlheuristik abhängt. Die Wahl der Heuristik beeinflußt jedoch signifikant die Größe - vor allem des ungeprunten - Klassifikationsbaumes. Die Interpretierbarkeit eines induzierten Klassifikationsbaumes nimmt direkt mit der Größe des Baumes (Tiefe der einzelnen Äste und Zahl der Äste) ab (Quinlan, 1986; Mingers, 1986). Aus diesem Grunde wurde neben dem von Quinlan in der ursprünglichen Fas-

sung des ID3-Algorithmus verwendeten Entropie-Kriterium in der vorliegenden Implementation auch noch das GINI-Kriterium als alternative Auswahlheuristik vorgesehen.

2.3. Methoden zur Gütebestimmung der induzierten Regeln

Hold-Out-Methode:

Bei dieser Methode wird der ursprüngliche Datensatz X, welcher N(X) Objekte enthält, zufällig in zwei disjunkte Teilmengen X' und X" von Objekten aufgeteilt, wobei X' insgesamt N(X') = a und X" insgesamt N(X") = 1-a Objekte der im im gesamten Datensatz X enthaltenen N(X) Objekte umfaßt. Der Wert von a liegt dabei zwischen 0.0 und 1.0, wobei meist 0.5 <= a <= 0.8 gewählt wird. X' wird als Trainingsdatensatz verwendet und X" als Verifikationsdatensatz. Mit dem Induktionsalgorithmus wird nun aus dem Trainingsdatensatz X' ein Regelsatz induziert. Zur Verifikation wird der Regelsatz dann auf den Verifikationsdatensatz X" angewandt. Das Verhältnis der dabei erhaltenen Zahl der Fehlklassifikationen zu der Gesamtzahl aller Objekte in X" ist ein Maß für die Güte des induzierten Regelsatzes.

Der wesentliche Nachteil der Hold-Out-Methode (HO-Methode) liegt in der ineffizienten Nutzung der in den verfügbaren Beispieldaten enthaltenen Information: je nach Aufteilungsverhältnis a werden nur a*100% der verfügbaren Daten zur Regelinduktion verwendet. Darüberhinaus besteht eine gewisse Abhängigkeit von der zufälligen Partition von X in X' und X". Je kleiner die Gesamtzahl von Objekten in X ist desto größer wird der resultierende Stichprobenfehler bei der Fehlerabschätzung sein. Erst bei einer sehr großen Gesamtzahl von Objekten in X wird der Stichprobenfehler vernachläßigbar sein. Die mit dieser Methode resultierende Fehlerabschätzung ist pessimistisch.

Leave-One-Out-Methode:

Eine effizientere Nutzung der im Gesamtdatensatz enthaltenen Information erhält man mit Hilfe der Leave-One-Out-Methode (LOO). Dabei wird wie folgt vorgegangen:

LOO-Schritt 1: Für alle Objekte O_i, i = 1,2,...,N des verfügbaren Datensatzes X führe die LOO-Schritte 2 bis 4 aus:

LOO-Schritt 2: Entferne das i-te Objekt O_i aus X, dies ergibt den i-ten Trainingsdatensatz $X_i = X \backslash \{O_i\}$. Setze den Fehlerindikator $f_i = 0$.

LOO-Schritt 3: Führe die Regelinduktion mit dem Trainingsdatensatz X_i durch.

LOO-Schritt 4: Teste den aus X_i induzierten Regelsatz mit dem Objekt O_i. Falls O_i falsch klassifiziert wird, setze den Fehlerindikator $f_i = 1$. Falls i < N gehe nach LOO-Schritt 1 (nächstes i).

LOO-Schritt 5: Als Schätzer für die Fehlerwahrscheinlichkeit wird der Quotient aus der Summe der f_i über i = 1,...,N und der Gesamtzahl der Objekte N verwendet.

Die LOO-Methode bringt einen relativ großen Berechnungsaufwand mit sich und ist daher vor allem für relativ kleine Datensätze anwendbar. Im Gegensatz zur HO-Methode nutzt sie aber die im Datensatz enthaltene Information effizienter. Aus diesem Grunde wurde die LOO-Methode im vorliegenden Fall zur Abschätzung des mit einem Regelsatz verbundenen Fehlers angewandt.

3. Ergebnisse

Aus dem Trainingsdatensatz wurden sowohl mit der Entropie- als auch mit der GINI-Heuristik die Induktion jeweils mit und ohne Pruning durchgeführt.

In den Abb.2a bis Abb.2d sind die mit ID3/C4 induzierten Entscheidungsbäume abgebildet. Die Blätter des Baumes repräsentieren die jeweiligen Klassen. d.h. der Nutzungstypen (1 bis 5). Dabei sind bei den einzelnen Regeln hinter der jeweiligen Klasse in Klammern die Anzahl n_r der durch diese Regel insgesamt klassifizierten Objekte bzw. die Anzahl n_f der durch die Regel falsch klassifizierten Objekte angegeben (n_r / n_f).

Tab.1 faßt die Ergebnisse aus den Abb.2a bis 2d zusammen.

Entscheidungsbaum:	Entropie ungeprunt	Entropie geprunt	GINI ungeprunt	GINI geprunt
Anzahl der Attribute	8	6	8	6
Anzahl der Regeln	14	10	13	9
Anzahl der Terme	64	37	51	31
Anzahl der Fehlklassifikationen	0	3	0	4

Tab.1: Charkteristische Kenngrößen der induzierten Entscheidungsbäume

```
if Lf < 69.2
            & SO4 < 68.75
                & Ba < 266.5
1.:                 & Na < 40.35                    then NT = 4 (9/0)
2.:                 & Na >= 40.35                   then NT = 2 (1/0)
3.:             & Ba >= 266.5                       then NT = 2 (2/0)
4.:         & SO4 >= 68.75                          then NT = 1 (10/0)
if Lf >= 69.2
            & Ca < 164.35
                & NO3 < 33.0
                    & Cl < 27.2
                        & SO4 < 93.8
5.:                         & Lf < 73.85            then NT = 2 (1/0)
6.:                         & Lf >= 73.85           then NT = 3 (2/0)
7.:                     & SO4 >= 93.8               then NT = 1 (2/0)
                    & Cl >= 27.2
8.:                     & Lf < 75.45                then NT = 1 (2/0)
                        & Lf >= 75.45
9.:                         & O2 < 5.35             then NT = 2 (17/0)
                            & O2 >= 5.35
10.:                            & Lf < 79.25        then NT = 1 (1/0)
11.:                            & Lf >= 79.25       then NT = 2 (1/0)
12.:            & NO3 >= 33.0                       then NT = 3 (7/0)
            & Ca >= 164.35
13.:            & SO4 < 96.8                        then NT = 3 (2/0)
14.:            & SO4 >= 96.8                       then NT = 5 (6/0)
```

Abb.2a: Induktion mit Entropie-Heuristik (ohne Pruning)

```
if Lf < 69.2
            & SO4 < 68.75
1.:             & Ba < 266.5                        then NT = 4 (10/1)
2.:             & Ba >= 266.5                       then NT = 2 (2/0)
3.:         & SO4 >= 68.75                          then NT = 1 (10/0)
if Lf >= 69.2
            & Ca < 164.35
                & NO3 < 33.0
                    & Cl < 27.2
4.:                     & SO4 < 93.8                then NT = 3 (3/1)
5.:                     & SO4 >= 93.8               then NT = 1 (2/0)
                    & Cl >= 27.2
6.:                     & Lf < 75.45                then NT = 1 (2/0)
7.:                     & Lf >= 75.45               then NT = 2 (19/1)
8.:             & NO3 >= 33.0                       then NT = 3 (7/0)
            & Ca >= 164.35
9.:             & SO4 < 96.8                        then NT = 3 (2/0)
10.:            & SO4 >= 96.8                       then NT = 5 (6/0)
```

Abb.2b: Induktion mit Entropie-Heuristik (mit Pruning)

```
if Lf < 75.55
            & SO4 < 71.95
1.:                 & Lf < 60.75                          then NT = 4 (7/0)
                    & Lf >= 60.75
2.:                         & Mn < 0.009                 then NT = 3 (2/0)
                            & Mn >= 0.009
3.:                                 & Fe < 1.84          then NT = 2 (4/0)
4.:                                 & Fe >= 1.84         then NT = 4 (2/0)
            & SO4 >= 71.95
5.:                 & NO3 < 1.1                          then NT = 3 (1/0)
6.:                 & NO3 >= 1.1                         then NT = 1 (13/0)
if Lf >= 75.55
            & Ca < 164.35
                    & NO3 < 25.8
7.:                         & O2 < 4.75                  then NT = 2 (16/0)
                            & O2 >= 4.75
8.:                                 & Lf < 79.45         then NT = 1 (2/0)
9.:                                 & Lf >= 79.45        then NT = 2 (1/0)
                    & NO3 >= 25.8
10.:                        & KS43 < 4.65                then NT = 2 (1/0)
11.:                        & KS43 >= 4.65               then NT = 3 (6/0)
            & Ca >= 164.35
12.:                & SO4 < 96.8                         then NT = 3 (2/0)
13.:                & SO4 >= 96.8                        then NT = 5 (6/0)
```

Abb.2c: Induktion mit GINI-Heuristik (ohne Pruning)

```
if Lf < 75.55
            & SO4 < 71.95
1.:                 & Lf < 60.75                          then NT = 4 (7/0)
                    & Lf >= 60.75
2.:                         & Mn < 0.009                 then NT = 3 (2/0)
                            & Mn >= 0.009
3.:                                 & Fe < 1.84          then NT = 2 (4/0)
4.:                                 & Fe >= 1.84         then NT = 4 (2/0)
5.:         & SO4 >= 71.95                               then NT = 1 (14/1)
if Lf >= 75.55
            & Ca < 164.35
6.:                 & NO3 < 25.8                         then NT = 2 (19/2)
7.:                 & NO3 >= 25.8                        then NT = 3 (7/1)
            & Ca >= 164.35
8.:                 & SO4 < 96.8                         then NT = 3 (2/0)
9.:                 & SO4 >= 96.8                        then NT = 5 (6/0)
```

Abb.2d: Induktion mit GINI-Heuristik (mit Pruning)

Eine formale Interpretation der induzierten Entscheidungsbäume zeigt, daß bereits im ungeprunten Baum mit fünf Regeln (bei dem mit dem Entropie-Kriterium erzeugten Baum: Regeln 1, 4, 9, 12 und 14 (Abb.2a); beim dem mit dem GINI-Kriterium erzeugten Baum: Regeln 1, 6, 7, 11 und 13 (Abb.2c)) bereits 78% bzw. 76% aller Objekte der Trainingsmenge richtig klassifiziert werden. Das bedeutet, daß für jede Nutzungsklasse genau eine Klassifikationsregel benötigt wird. Die restlichen Regeln (Entropie: 9 Regeln; GINI: 8 Regeln) werden nur zur Klassifikation der restlichen 22% bzw. 24% der Objekte benötigt. Dies deutet darauf hin, daß nur wenige (sprich 5) Regeln notwendig sind, um über 75% der in der Trainingsmenge enthaltenen und für eine Klassifikation der Nutzungstypen notwendigen Information zu erfassen. Dies kann so interpretiert werden, daß die Regeln eine relativ hohe Verallgemeinerung und damit eine kompakte Repräsentation der relevanten Information darstellen.

Unabhängig von der gewählten Auswahlheuristik werden die 5 Nutzungstypen (Klassen) im wesentlichen durch die 4 Attribute Lf, SO_4, Ca und NO_3 charakterisiert. Beide Heuristiken identifizieren Lf (Leitfähigkeit) als das Attribut mit der höchsten Trennschärfe, gefolgt von SO_4 (Sulfat) bzw. Ca (Calcium) und schließlich NO_3 (Nitrat). Die den jeweiligen Nutzungstypen zugeordneten Wertebereiche der Attribute sind im Hinblick auf die Nutzungscharakteristik schlüssig und fachlich erklärbar (Hiessl, 1992).

Mit Hilfe der Leave-One-Out-Methode wurden dann sowohl mit der Entropie- als auch mit der GINI-Heuristik Fehlerabschätzungen durchgeführt und damit die Robustheit des jeweiligen, aus der gesamten Trainingsmenge ermittelten, ungeprunten (Abb.2a und Abb.2c) sowie der geprunten (Abb.2b und Abb.2d) Entscheidungsbäume überprüft. Die Ergebnisse der Anwendung der LOO-Methode sind in Tab.2 zusammengestellt. Sie werden am Beispiel des GINI-Kriteriums (Fall: ungeprunt) erläutert. "A" besagt: in 39 Fällen klassifiziert der aus den "Restbeispielen" induzierte Regelsatz das weggelassene (ungesehene) Objekt richtig. "A" stellt somit ein Maß für die Allgemeingültigkeit der Regelmenge dar. "B" besagt das Gegenteil: in 24 Fällen konnte der jeweils induzierte Regelsatz das unbekannte (weggelassene) Objekt nicht richtig klassifizieren. In "A" und "B" manifestiert sich die Genauigkeit der Klassifikation des ungeprunten, d.h. exakten Regelbaums. "C" und "D" sind Teilmengen von "B" zur Prüfung der Wirksamkeit des Baumbeschneidungsverfahrens. Die 4 Objekte in "D" wurden sowohl von der LOO-Methode als auch durch das Pruning als verrauscht identifiziert. "C" stellt ein Maß für die Stärke der durchgeführten Baumbeschneidung dar. In 20 von 24 Fällen wurde bei GINI auf ein Pruning verzichtet. Das Ergebnis von E1 und E2 ist dahingehend zu interpretieren, daß das GINI-Kriterium die in den Beispielen enthaltene Informationen vollständig nutzt. E1 besagt, daß in 6 Fällen das Weglassen eines Objektes zu einer umfangreicheren Regelmenge führte. E2 dokumentiert, ob weggelassene Objekte zu einer kleineren Regelmenge führten. Die LOO-Methode konnte für den "GINI-Regelsatz" keine Redundanzen feststellen.

Beschreibung	Entropie ungeprunt	GINI ungeprunt	Entropie geprunt	GINI geprunt
A.] Bei LOO ausgelassene Objekte werden im vollständigen-und im LOO-Baum richtig klassifiziert	41	39	37	39
B.] Bei LOO ausgelassene Objekte werden im vollständigen Baum richtig und im LOO-Baum falsch klassifiziert	22	24	26	24
C.] Bei LOO ausgelassene Objekte werden im geprunten richtig und im LOO-Baum falsch klassifiziert	19	20	23	20
D.] Ausgelassene Objekte werden sowohl im vollständigen Baum geprunt und falsch klassifiziert als auch im LOO-Baum falsch klassifiziert	3	4	3	4
E1.] LOO-Baum umfangreicher (bezüglich Anzahl der Regeln/Terme/verwendeten Attribute) als der vollständige Baum	0	6	19	4
E2.] LOO-Baum kleiner (bezüglich Anzahl der Regeln/Terme/verwendete Attribute) als der vollständige Baum	15	0	17	11

Tab.2: Ergebnisse der Anwendung der Leave-One-Out-Methode und Vergleich mit der Induktion anhand des vollständigen Datensatzes.

Eine entsprechende Interpretation kann auch für den Fall "geprunt" erfolgen, bei dem in jedem LOO-Baum ein Pruning-Zyklus durchgeführt wurde. Durch die LOO-Methode ergaben sich folgende Fehlerabschätzungen für die vier induzierten Klassifikationsbäume:

Klassifikationsbaum:	Entropie ungeprunt	GINI ungeprunt	Entropie geprunt	GINI geprunt
LOO-Fehler:	35%	38%	41%	38%

Der mit der LOO-Methode ermittelte relativ hohe Fehler der verschiedenen Klassifikationsbäume deutet auf einen hohen Rauschanteil in den Daten hin.

Es ist festzuhalten, daß bei Anwendung der LOO-Methode zur Schätzfehlerermittlung im Falle der Entropie-Heuristik eine Verschlechterung eintritt, wenn der Regelbaum geprunt wird. Der mittels LOO-Methode abgeschätzte Fehler der GINI-Heuristik ist unabhängig davon, ob der

Baum geprunt wird oder nicht. Daher kann vermutet werden, daß ein relatives Konzentrationsmaß oder auch Disparitätsmaß, wie es das GINI-Kriterium darstellt, im Falle von verrauschten Daten, eine bessere Heuristik im Vergleich zu einem absoluten Konzentrationsmaß ist, wie es bspw. das Entropie-Kriterium darstellt. Diese Vermutung konnte an dieser Stelle jedoch nicht weiter untersucht werden.

Aufgrund der Untersuchungsergebnisse mit diesem Datensatz kann festgehalten werden, daß die GINI-Heuristik kompaktere (im Sinne von weniger Regeln und weniger Termen) und auch stabilere Regelsätze liefert. Der LOO-Fehler ist im Gegensatz zum Entropie-Kriterium unabhängig davon, ob die LOO-Prozedur mittels ungeprunten oder geprunten Bäumen durchgeführt wird. Ferner spricht für das GINI-Kriterium, daß der LOO-Schätzfehler bei Verwendung geprunter Regelbäume um 8% geringer ist als der entsprechende Fehler des mittels Entropie-Kriteriums induzierten Baums.

4. Zusammenfassung

Die Grundwasserqualität wird mit Hilfe von Beobachtungsbrunnen überwacht. Die chemisch-biologische Charakteristik des Grundwassers in einem solchen Brunnen wird vor allem von der Landnutzung im Eintragsbereich des Brunnens bestimmt. Die Kenntnis dieses Zusammenhanges ist daher für den Grundwasserschutz von großer Bedeutung.

In dem Beitrag wurde der ID3/C4-Algorithmus angewandt, um für einen Grundwasserdatensatz (63 Objekte = Brunnen, 22 stetige Attribute = Grundwasserbeschaffenheitsparameter und 5 Klassen = Landnutzungstypen) diese Beziehung abzuleiten. Es wurden 2 Varianten des Algorithmus mit unterschiedlichen heuristischen Auswahlkriterien (Entropie- und GINI-Kriterium) verglichen. Aufgrund des geringen Umfanges des Datensatzes konnte die üblicherweise eingesetzte "Holdout"-Methode zu Verifikation und Gütebewertung der mit der jeweiligen Variante induzierten Regelsätze bei der Klassifikation neuer, d.h. nicht zum Lernen verwandter Daten, nicht angewandt werden. Um die in dem Datensatz enthaltene Information effizient zu nutzen, wurde stattdessen die "Leave-One-Out"-Methode verwendet.

Es zeigt sich, daß das Gini-Kriterium sowohl stabilere als auch kompaktere Regelsätze induziert. Die "Leave-One-Out"-Methode bestätigt die Anwendung des "Pessimistic-Pruning" vor allem bei Einsatz der GINI-Heuristik. Beim Pruning werden i.d.R. jene Regeln entfernt, welche auch im Falle des Auslassens der von ihnen abgedeckten Beispiele im Zuge der "Leave-One-Out"-Methode eine Kompaktierung des Regelsatzes zur Folge haben.

Insgesamt hat sich in der hier vorgestellten ersten Anwendung eines Regelinduktionsverfahrens zur Klassifikation der Eintragsbereiche von Grundwasserbeobachtungsbrunnen anhand der

vorherrschenden Landnutzungen gezeigt, daß mit Hilfe der Regelinduktion interessante und fachlich interpretierbare Ergebnisse erzielt werden können.

5. Literatur

Breiman, L. et al. (1984): Classification and Regressions Trees. Wadsworth International Group, Belmont, CA.

Grimm-Strele, J. et al. (1991): Modellhafte Einrichtung eines Grundwassergüte-Meßnetzes in einer ausgewählten Region ("Modellmeßnetz Karlsruhe"). Abschlußbericht zum Forschungsprojekt "Wasser 102 04 214" an das Umweltbundesamt, Berlin.

Hiessl, H. (1992): Regelinduktionsverfahren als Mittel zur Klassifikation von Landnutzungstypen aus Grundwassergütedaten. S.108-118. In: Grimm-Strele, J. et al. (Hrsg., 1992): Grundwasserüberwachungsprogramm - Pilotprojekt Karlsruhe, Landesanstalt für Umweltschutz Baden-Württemberg, Reihe LfU-Wasser, Heft 6, Karlsruhe, 1992.

Marfels, Ch. (1977): Erfassung und Darstellung industrieller Konzentration. Baden-Baden, 1977.

Mingers, J. (1986): Inducing Rules for Expert Systems Statistical Aspects. The Professional Statistican, 5, pp.19.

Mingers, J. (1989): An Empirical Comparison of Selection Measures for Decision-Tree Induction. KI Academic Publishers, Machine Learning 3, pp.319.

Quinlan, J.R. (1979): Discovering Rules by Induction from Large Collections of Examples. In: D. Michie (ed., 1979): Expert Systems in the Micro Electronic Age. Edinburgh University Press, 1979.

Quinlan, J.R. (1983): Learning Efficient Classification Procedures and Their Application to Chess Endgames. pp.463-482. In: R.S. Michalski; J.G. Carbonell; T.M. Mitchell (eds., 1983): Machine Learning: An Artificial Intelligence Approach. Springer Verlag, Berlin.

Quinlan, J.R. (1986): Induction of Decision Trees. Machine Learning 1, pp.81-106, 1986.

Quinlan, J.R. (1987): Simplifying Decision Trees. Int.J.Man-Machine Studies, Vol.27, pp.221-234, 1987.