Perlen der theoretischen Informatik

Seminar am Lehrstuhl für theoretische Informatik der Universität GH Paderborn

Thema: Algorithmisches Lernen mit der Methode der induktiven Inferenz

Anja Austermann

I. Was ist induktive Inferenz?

Damit ein Computer in der Lage ist, intelligentes Verhalten zu zeigen, muß er über eine Wissensbasis verfügen, aus der er Schlußfolgerungen ziehen kann. In den klassischen, wissensbasierten Systemen wird das Problem, eine Wissensbasis aufzubauen, in der Regel gelöst, indem ein menschlicher Experte explizit eine Menge von Fakten und Regeln angibt, die das System als Grundlage für intelligentes Verhalten verwenden kann. Diese Art, einen Computer mit Wissen auszustatten, funktioniert insbesondere in Spezialgebieten wie beispielsweise bei der Fehlerdiagnose von technischen Systemen recht gut. Allerdings kann sich ein solches System ohne Hilfe eines menschlichen Experten kein zusätzliches Wissen aneignen. An dieser Stelle kommt das algorithmische Lernen ins Spiel. Hier soll eine spezielle Methode des algorithmischen Lernens, die induktive Inferenz nach dem Paradigma von Gold, näher vorgestellt werden.

Eine typische Anwendung der induktiven Inferenz ist aus Intelligenztests bekannt, in denen der Proband aufgefordert wird, eine Zahlenreihe sinnvoll fortzusetzen. Die nächste Zahl in der Reihe

0, 2, 4, 6, ..

ist natürlich die 8, da zu jedem Wert die Zahl 2 addiert wird. Was es uns Menschen so einfach macht, diesen Zusammenhang zu erkennen, ist unsere Fähigkeit, induktiv zu schlußfolgern und diese Fähigkeit wird mit der Methode der induktiven Inferenz am Computer nachmodelliert.
Induktive Inferenz bedeutet generell, aus einer Reihe von gegebenen Beispielen eine gültige Regel zu schlußfolgern, nach der die Beispielsequenz gebildet werden kann. Eine Regel wird als gültig akzeptiert, wenn das System mit Hilfe dieser Regel in der Lage ist, alle zukünftigen Beispiele - also bei einer Zahlenreihe alle folgenden Zahlen - korrekt vorherzusagen. Die Anwendung dieses Prinzips ist natürlich nicht nur für die Vervollständigung von Zahlenreihen zu gebrauchen, sondern dient beispielsweise auch dazu, ganze Turing-Maschinen oder Automaten aus Beispielen von Wörtern zu konstruieren, die von ihnen akzeptiert beziehungsweise nicht akzeptiert werden, oder aus Traces eines Programmablaufs das zugrunde liegende Programm zu rekonstruieren.
Hier stellt sich die Frage, über welche Formen von Regeln generell Aussagen mit Hilfe des Prinzips der induktiven Inferenz gemacht werden können. Ein intuitives Beispiel für eine Regel, die von einem Inferenzsystem nicht korrekt erkannt werden kann, ist die folgende: "Jede Zahl, die ausgegeben wird, ist genau um eins größer, als die Vorhersage des Inferenzsystems" Es wird sofort klar, wieso eine korrekte Vorhersage dieser Zahlenreihe nicht möglich ist, und daß die Vorhersage immer um genau -1 von der tatsächlichen Aussage abweicht. Damit ein System eine Regel korrekt vorhersagen kann, darf sie sich nicht auf das Inferenzsystem und dessen Vorhersagen beziehen. Außerdem müssen alle Parameter, auf die sich eine Regel bezieht, für das System sichtbar sein.

II. Spezifikation von Inferenzproblemen

Ein Inferenzproblem ist durch folgende fünf Angaben vollständig spezifiziert:
(1) die Klasse R der für die Identifikation in Betracht kommenden Regeln
(2) die Klasse H der als Hypothesen in Betracht kommenden Beschreibungen von Regeln
(3) eine Sequenz B von Beispielen
(4) die zu verwendenden Inferenzmethoden
(5) und einer Reihe von Kriterien wann ein Inferenzversuch als erfolgreich bezeichnet wird

zu 1): Eine Regel ist normalerweise entweder eine Sprache über einem beliebigen Alphabet oder eine ein- oder mehrdimensionale Funktion, die Ausgangswerten aus einem Definitionsbereich, wie beispielsweise Positionen in einer Zahlenreihe, Werte aus einem Wertebereich zuordnet.

Im ersten Beispiel sind sowohl der Definitionsbereich als auch der Wertebereich die positiven ganzen Zahlen, falls nicht auch Zahlenreihen, die aus Brüchen oder negativen Zahlen bestehen, zugelassen sind. Als Regelklasse kann man hier die Klasse der Polynome beliebigen Grades über den ganzen Zahlen auffassen.

zu 2): Die Klasse H der als Hypothesen in Betracht kommenden Beschreibungen unterscheidet sich von R in sofern, daß eine Regel mehrere unterschiedliche Beschreibungen haben kann. Für jede Regel

aus R muß allerdings mindestens eine Beschreibung in H existieren, da ein Inferenzsystem eine Regel, die keine Beschreibung in H hat, nicht erfolgreich identifizieren kann.

Der Hypothesenraum für Zahlenreihen ist häufig die Menge der Polynome beliebigen Grades in einer Variable. Die Variable steht für die für die Position einer Zahl in der Reihe. Außerdem wird der Typ der Koeffizienten auf ganze Zahlen eingeschränkt. Die Zahlenreihe aus unserem Beispiel hat folgende Beschreibung:

$f(n) = 2n$

Eine andere mögliche Beschreibung ist folgende rekursive Funktion:

$f(0) = 0$
$f(n+1) = f(n) + 2$

Es können also durchaus ganz unterschiedliche Hypothesen die gleiche Regel beschreiben.

zu 3): Ein Inferenzsystem schlußfolgert anhand von Beispielen. In erster Linie unterscheidet man hier, ob ausschließlich Positivbeispiele angegeben werden, oder ob zusätzlich Negativbeispiele zugelassen sind. Letztere Variante ist mächtiger, da hier gezielt Regeln ausgeschlossen werden können, während bei Positivbeispielen immer eine Zusatzbedingung zum Schlußfolgern erforderlich ist, wie beispielsweise, daß die zuerst gefundene, die einfachste oder die genaueste Regel ausgewählt werden soll.

Beispielsequenzen müssen einer bestimmten Darstellungsform genügen, die zusammen mit dem Inferenzsystem spezifiziert wird. Für eine Zahlenreihe könnte das zum Beispiel <0, f(0)>, <1, f(1)>, ...<n, f(n)> sein, wobei die erste Zahl immer für die Position in der Zahlenreihe und die zweite für den eigentlichen Wert steht. Da die Positionen fortlaufend durchnumeriert sind, wäre aber auch f(0), f(1), ..., f(n) eine eindeutige und zulässige Darstellungsform. Hierbei müssen alle Beispiele Elemente des Graphen von f sein. Die beiden Darstellungsformen oben bestehen ausschließlich aus positiven Beispielen. Eine Beispielsequenz, die auch negative Beispiele beinhaltet könnte beispielsweise folgende sein: <aabb, yes>, <aaab, no>, <ab, yes>, <aaabbb, yes>, <abbb, no>. Sie repräsentiert die Sprache über dem Alphabet {a,b}, bei der Alle Wörter die Form a^nb^n haben.

zu 4): Als Klasse der möglichen Inferenzmethoden ist es im Regelfall sinnvoll, alle von einem Computer berechenbaren Methoden auszuwählen, die eine Folge von Zahlen oder Worten über einem Alphabet akzeptieren, endlich lange rechnen und dann eine Ausgabe passend zur Eingabe produzieren.
Eine Inferenzmethode für die Berechnung von Zahlenreihen akzeptiert beispielsweise eine Sequenz von ganzen Zahlen und gibt, nachdem sie endlich lange gerechnet hat, eine ganze Zahl aus.

zu 5): Damit ein Inferenzsystem in der Lage ist, irgendwann eine Schlußfolgerung auszugeben, sind Kriterien notwendig, die angeben, wann eine Identifikation als erfolgreich anerkannt werden soll. Ein Kriterium ist hier die Identifikation im Limes, die im nächsten Abschnitt genauer betrachtet wird, aber es gibt auch andere Kriterien, die sich von der Identifikation im Limes in Gültigkeitsbereich, Stärke und Menge der inferierbaren Funktionen unterscheiden.

III. Identifikationsmethoden und Vorhersage

III.1. Identifikation im Limes

Nach dem Paradigma von Gold funktioniert induktive Inferenz nach der Methode der Identifikation im Limes. Identifikation im Limes betrachtet die induktive Inferenz als einen unendlichen Prozess und geht davon aus, daß eine unendliche Folge von Beispielen zur Verfügung steht. Anhand dieser Beispiele wird in jedem Schritt eine passende Hypothese aus H ausgewählt. Als Kriterium für den Erfolg eines Inferenzvorgangs wird das Konvergenzverhalten der Inferenzmethode verwendet. Eine Inferenzmethode konvergiert bei einer gegebenen Beispielsequenz gegen eine bestimmte Hypothese h, wenn sie entweder endlich ist und h die letzte von ihr berechnete Hypothese, oder unendlich, wobei die Hypothese h nach einer endlichen Zeit erreicht wird, und sich nicht weiter ändert und mit

ihrer Hilfe jedes folgende Beispiel korrekt vorhergesagt werden kann. Da theoretisch nach einer beliebigen Anzahl von Beispielen eins auftauchen kann, das der bisherigen Hypothese widerspricht, kann die Konvergenz dieser Methode allerdings zu keinem Zeitpunkt definitiv festgestellt werden.
Die Klasse aller Mengen von totalen rekursiven Funktionen, die von einer beliebigen Identifikationsmethode identifiziert werden können, wird im Folgenden als EX bezeichnet, wobei die Abkürzung EX für das englische Wort "explanatory" steht.
Eine Unterklasse von EX ist die Klasse BC. Sie beinhaltet die Mengen von totalen, rekursiven Funktionen, die durch BC-Identifikation identifiziert werden können. BC steht hierbei für "Behaviorally Correct", was soviel bedeutet wie "vom Verhalten her korrekt" und BC-Identifikation erlaubt, daß die Hypothese sich unendlich lange ändert und die Identifikationsmethode also nie konvergiert, solange mit ihrer Hilfe trotzdem korrekte Vorhersagen getroffen werden können. Das Verfahren der BC-Identifikation wird auch als Übereinstimmung im Limes bezeichnet und ist nicht mit der Identifikation im Limes zu verwechseln, bei der es notwendig ist, daß sich die Hypothese nach endlicher Zeit nicht mehr ändert.

III.2. Identifikation durch Aufzählung

Die Methode der Identifikation durch Aufzählung basiert auf dem systematischen Durchsuchen aller für eine Identifikation in Frage kommenden Regeln, bis eine davon mit allen bisher angegebenen Beispielen übereinstimmt. Damit die Identifikation durch Aufzählung berechenbar ist, muß die Menge der Regeln beziehungsweise deren Beschreibungen aufzählbar sein, und es muß außerdem möglich sein, eine Ordnung auf dieser Menge zu definieren. Wenn diese beiden Bedingungen auf ein Identifikationsproblem zutreffen, so kann eine geordnete Liste der zu durchsuchenden Hypothesen erzeugt werden. Die erste Hypothese, die beim Durchsuchen der Liste mit den bisherigen Beispielen übereinstimmt, wird als aktuelle Hypothese ausgewählt. Falls sichergestellt ist, daß

a) eine korrekte Hypothese immer zu den angegebenen Beispielen kompatibel ist und
b) jede inkorrekte Hypothese mit einer ausreichend großen Menge von Beispielen und allen größeren Mengen inkompatibel ist

so ist eine korrekte Identifikation im Limes gewährleistet. Obwohl die Identifikation durch Aufzählung ein sehr universelles und mächtiges Verfahren ist, gestaltet sich die praktische Verwendung eher schwierig, da die Anzahl der zu durchsuchenden Regeln exponentiell mit der Länge ihrer Beschreibungen wächst. Deswegen werden in der Praxis häufig Einschränkungen dieses Verfahrens angewandt, die die Anzahl der zu durchsuchenden Regeln verringern.

Vorhersage statt Identifikation

Häufig ist nicht die Regel nach der eine Sequenz von Beispielen erzeugt wird von Interesse, sondern es soll eine korrekte Vorhersage des nächsten Beispiels getroffen werden. In diesem Fall werden statt der oben beschriebenen Identifikationsmethoden Vorhersagemethoden eingesetzt. Eine Vorhersagemethode akzeptiert als Eingabe eine Sequenz von n Werten und verwendet diese, um den Wert an der Stelle n+1 vorherzusagen, sofern dies möglich ist. Für den Erfolg einer Vorhersagemethode gibt es drei verschiedene Definitionen, die sich darin unterscheiden, ob die Vorhersagemethode in jedem Fall konvergieren muß, oder ob Beispiele erlaubt sind, für die die Vorhersagemethode nicht konvergiert.
Die stärksten Einschränkungen treten bei der NV-Vorhersage auf. Eine Inferenzmethode M trifft genau dann eine NV-Vorhersage über eine totale, rekursive Funktion, wenn $M(x_1, x_2,...,x_n)$ für alle endlichen Sequenzen von Werten $x_1, x_2,...,x_n$ definiert ist, und weiterhin $M(f(0), f(1),...,f(n-1))=f(n)$ für alle bis auf endlich viele n gilt. Die Methode m muß also aus beliebigen Werten f(0) - f(n-1) korrekt den Wert f(n) vorhersagen.
Wenn U eine Menge totaler, rekursiver Funktionen ist, dann triff M genau dann eine NV-Vorhersage über U, wenn M jede Funktion f aus U NV-vorhersagt. Im Unterschied zur NV' und NV''-Vorhersage, muß M im Fall einer NV-Vorhersage auf allen endlichen Sequenzen konvergieren.
Etwas weniger strikt geht die Methode der NV'-Vorhersage vor. Hier muß $M(x_1,x_2,...,x_n)$ nicht für beliebige x_i definiert sein, sondern nur für solche, die Funktionswerte von f an einer beliebigen Stelle sind. Es muß also M(f(0), f(1), ..., f(n-1)) für alle n definiert sein. Außerdem muß wie oben an allen Stellen, wo M definiert ist, gelten, daß M(f(0), f(1),...,f(n-1)) = f(n). M konvergiert bei der NV'-Vorhersage auf allen Anfangssequenzen von Funktionen, deren Werte korrekt vorhergesagt werden.

Im Fall der NV''-Vorhersage fällt die Bedingung, für welche x_1 bis x_n $M(x_1, x_2, ...,x_n)$ definiert sein muß, weg, und es ist lediglich verlangt, daß M(f(0), f(1),...,f(n-1)) an allen bis auf endlich vielen Stellen definiert und an diesen Stellen gleich dem Wert f(n) ist. Hier ist es möglich, daß M auf endlich vielen Anfangssequenzen nicht konvergiert, obwohl für deren Funktionen korrekte Vorhersagen getroffen werden können
Die Klassen NV, NV' und NV'' beinhalten jeweils alle totalen, rekursiven Funktionen, deren Funktionswerte von einer NV- beziehungsweise NV' oder NV''-Methode korrekt vorhergesagt werden. Hierbei ist zu beachten, daß NV eine echte Unterklasse von NV' ist, welche wiederum eine echte Unterklasse von NV'' darstellt. NV'' entspricht der weiter oben beschriebenen Klasse BC.
Obwohl es auf den ersten Blick so aussehen mag, als wären die Anwendungsgebiete für Vorhersage und Identifikation die Gleichen, stellt sich bei näherer Betrachtung heraus, daß zwar immer dann, wenn Vorhersage möglich ist, auch Identifikation möglich ist, das Umgekehrte aber nur dann gilt, wenn ausschließlich totale rekursive Funktionen als Hypothesen für die Identifikation in Frage kommen. Ist das der Fall, kann die Hypothese, die sich durch Identifikation ergeben hat, zur Vorhersage von Funktionen benutzt werden, indem sie auf die bisherigen Werte der Beispielsequenz angewandt wird, und das Ergebnis als Vorhersage des nächsten Wertes benutzt wird.

Aus einer NV'-Vorhersagemethode läßt sich umgekehrt wie folgt eine Identifikationsmethode konstruieren:
Sei A ein Algorithmus, der eine NV'-Vorhersagemethode berechnet, so kann daraus folgendermaßen eine Hypothese f erzeugt werden, wenn eine Sequenz $x_1,...,x_n$ von Beispielen vorgegeben ist:

$f(i) = x_i$ für alle x_i aus der Beispielsequenz
$f(i+1) = A(f(1), f(2),...,f(i))$ für alle $i >= n$

Die Hypothese, die von der Identifikationsmethode berechnet wird, ist also eine Funktion, bei der jeder Funktionswert als Ergebnis einer Vorhersage durch den Algorithmus A definiert wird. Hieraus ergibt sich eine korrekte Identifikation im Limes, sofern A im Limes korrekte Vorhersagen liefert.

IV. Vergleichbarkeit von Inferenzmethoden

Um Inferenzmethoden miteinander zu vergleichen, braucht man Kriterien zur Bewertung. Die drei Kriterien, die hierfür hauptsächlich herangezogen werden sind

(1) Qualität der Hypothesen
(2) Dateneffizienz
(3) die Anzahl verschiedener Hypothesen beziehungsweise die Anzahl von Fehlern und Änderungen der aktuellen Hypothese

zu 1) Für die Qualität der Hypothesen gibt es zwei grundsätzliche Kriterien: ihre Einfachheit sowie ihre Relation zu den angegebenen Beispieldaten. Beide Kriterien generieren eine Ordnungsrelation auf der Hypothesenmenge. Die Relation $g < h$ bedeutet in diesem Fall, daß g die Beispielssequenz S besser beschreibt, als h. Die Ordnungsrelation auf einer Hypothesenmenge kann von der Beispielsequenz S abhängig oder unabhängig sein. Sie ist unabhängig von der Beispielsequenz, genau dann, wenn g alle möglichen Beispielssequenzen, die zu beiden Hypothesen kompatibel sind, besser beschreibt als h, sobald g irgendeine Beispielsequenz besser beschreibt, als h. Alle Ordnungen die sich ausschließlich aus Eigenschaften der Hypothesen, also beispielsweise deren Länge oder Ordnung, berechnen lassen, sind unabhängig von der gegebenen Beispielsequenz.

Ordnung nach Einfachheit
Ein Beispiel für eine Ordnung, die unabhängig von der gegebenen Beispielsequenz ist, ist die Ordnung nach Einfachheit der Hypothese. Hier sind die Hypothesen die besten, die die geringste Komplexität aufweisen. Bei Polynomen sind das beispielsweise die mit dem niedrigsten Grad, oder bei Automaten die mit den wenigsten Zuständen. Die Methode der Identifikation durch Aufzählung kann verwendet werden, um die einfachste Hypothese zu ermitteln, die mit allen positiven Beispielen übereinstimmt und, falls Negativbeispiele vorhanden sind, mit keinem der Negativbeispiele übereinstimmt. Damit die Methode der Identifikation durch Aufzählung dies leisten kann, muß allerdings die Menge der zu

testenden Hypothesen nach Einfachheit geordnet sein, so daß die erste Hypothese, die mit den gegebenen Beispielen übereinstimmt, auch in jedem Fall die einfachste ist.

Ordnung nach Genauigkeit
Wenn eine Sprache ausschließlich mit Hilfe positiver Beispiele inferiert werden soll, so ist die Ordnung nach Genauigkeit ein wichtiges Kriterium. Hierbei werden Hypothesen nach der Menge der Wörter der Sprache L(h), also der Sprache, die von der Hypothese h generiert wird, geordnet. Ist L(g) eine Untermenge von L(h), so ist g eine bessere Hypothese, als h. Ist die Menge der zu testenden Hypothesen nach Genauigkeit geordnet, so wird bei der Identifikation durch Aufzählung immer die Hypothese ausgewählt, die alle Beispiele erzeugt, aber so wenig zusätzliche Wörter, wie möglich.

zu 2) Dateneffizienz ist ein Kriterium zur Bewertung von Inferenzmethoden, das eine Aussage darüber macht, wie viele Beispiele eine Inferenzmethode lesen muß, bis sie beim n-ten Beispiel zu einer endgültigen Hypothese gelangt, also wenn für alle i >= n gilt $h_i+1 = h_i$. Der Wert n wird auch Konvergenzpunkt genannt. Eine Inferenzmethode M_2 ist genau dann so dateneffizient wie eine Inferenzmethode M_1, wenn die beiden Methoden für alle möglichen zulässigen Beispielsequenzen die gleichen Regeln identifizieren und den der Konvergenzpunkt von M_2 nicht größer ist, als der von M_1.
Ist eine Inferenzmethode M_2 so dateneffizient wie M_1 und gilt außerdem, daß es mindestens eine Regel gibt, bei deren Identifikation M_2 einen kleineren Konvergenzpunkt hat, als M_1, so ist M_2 dateneffizienter, als M_1.
Existiert keine Inferenzmethode, die dateneffizienter ist, als die Methode M_2, so besitzt M_2 eine optimale Dateneffizienz.
Gold hat 1977 bewiesen, daß Identifikation durch Aufzählung eine Inferenzmethode mit optimaler Dateneffizienz ist.

zu 3) Ein anderes Kriterium zur Bewertung von Inferenzmethoden bezieht sich nicht auf die Anzahl der einzulesenden Beispiele, sondern darauf, wie viele Fehler eine Inferenzmethode macht, bevor sie den Konvergenzpunkt erreicht. Eng damit verbunden sind Kriterien, die sich darauf beziehen, wie viele verschiedene Hypothesen eine Inferenzmethode annimmt, bevor sie zum Konvergenzpunkt gelangt, oder wie häufig die aktuelle Hypothese gewechselt wird. Im ersten Moment mag es so aussehen, als würden beide Kriterien den gleichen Sachverhalt beschreiben, allerdings ist das nur scheinbar der Fall. Falls die Inferenzmethode konservativ ist, also die aktuelle Hypothese nur dann wechselt, wenn die bisherige Hypothese sich als falsch erweist, und wenn weiterhin gefordert ist, daß die aktuelle Hypothese ohne Ausnahmen mit allen Beispielen aus der angegebenen Beispielsequenz übereinstimmen muß, so ist die Anzahl der unterschiedlichen Hypothesen genau um eins größer, als die Anzahl der Hypothesenwechsel. Ist aber eine der beiden Voraussetzungen nicht gegeben, so kann es passieren, daß die Inferenzmethode später zu einer Hypothese zurückkehrt, die sie bereits verworfen hatte. In diesem Fall kann die Anzahl der unterschiedlichen Hypothesen auch sehr viel kleiner sein, als die Anzahl der Hypothesenwechsel.
Bei der Identifikation durch Aufzählung kann es im ungünstigsten Fall passieren, daß bei der Identifikation der n-ten Funktion in der Aufzählung n unterschiedliche Hypothesen angenommen werden und n-1 Hypothesenwechsel stattfinden. Barzin und Freivald haben im Jahr 1972 allerdings bewiesen, daß es eine Identifikationsmethode gibt, die nur n+o(log n) unterschiedliche Hypothesen benötigt und haben weiterhin gezeigt, daß log n - O(1) eine untere Schranke für dieses Problem ist.

V. Sinnvolle Beschränkungen für Inferenzmethoden

Es gibt verschiedene Beschränkungen für Inferenzmethoden. Diese Beschränkungen haben den Effekt, daß weniger Regelmengen identifiziert werden können, da bestimmte Inferenzmethoden, die den strengeren Anforderungen nicht genügen, nicht mehr verwendet werden dürfen

V.1. Konsistenz

Eine wichtige Einschränkung ist die Forderung nach Konsistenz. Eine konsistente Inferenzmethode gibt nur solche Hypothesen aus, die mit allen bisher gelesenen Beispielen übereinstimmen. Ein Beispiel für eine konsistente Methode ist die Identifikation durch Aufzählung. Es gibt aber beispielsweise auch stochastische Methoden, die das Abweichen einzelner Beispiele gestatten und somit nicht konsistent sind.

Es besteht ein Unterschied darin, ob Konsistenz für beliebige Beispielsequenzen notwendig ist, oder ob sie nur für Beispielsequenzen gefordert ist, die auch tatsächlich zu einer Regel passen und korrekt identifizierbar sind.
Die Klasse der rekursiven Funktionen, die durch im letzteren Sinne konsistente Inferenzmethoden identifiziert werden können, ist eine echte Unterklasse der Klasse EX aller identifizierbaren Funktionen. Es gibt also Funktionen, die zwar generell identifiziert werden können, für die es aber keine konsistente Identifikationsmethode gibt.

V.2. Endliche Identifizierbarkeit

Ein Problem bei der Identifikation im Limes ist die Tatsache, daß es unmöglich ist, Konvergenz festzustellen, so daß im Prinzip unendlich lange gerechnet werden muß, da prinzipiell zu jedem Zeitpunkt, zu dem sich die Inferenzmethode für eine Hypothese entscheidet, ein weiteres Beispiel die Hypothese widerlegen könnte.
Weil es in der Praxis aber in der Regel erforderlich ist, daß eine Inferenzmethode irgendwann zu einer Entscheidung gelangt, wird häufig Identifikation in endlicher Zeit gefordert. Freivald und Wiehagen definierten endliche Konvergenz wie folgt: Eine endliche oder unendliche Sequenz y_1, y_2, y_3... konvergiert endlich zu einem Wert t an der Stelle n, genau dann, wenn $y_n = t$ und n ist die kleinste, natürliche Zahl, so daß $y_n = y_n + 1$. Sobald ein Wert also in zwei aufeinander folgenden Positionen wiederholt wird, wird nach der Definition von Konvergenz ausgegangen.
Auch die Klasse der endlich identifizierbaren Funktionen ist eine echte Unterklasse der Klasse EX.

V.3. Zuverlässigkeit

Eine Inferenzmethode ist zuverlässig, wenn sie immer dann, wenn sie konvergiert, auch gegen eine korrekte Hypothese konvergiert und ansonsten keine Konvergenz eintritt. Wenn aus einer Sequenz also keine Regel identifizierbar ist, so muß die Inferenzmethode divergieren, also ihre Hypothese unendlich oft ändern. Das bedeutet, daß erfolgreiche Identifikation bei einer zuverlässigen Inferenzmethode genau dann auftritt, wenn die Anzahl der Hypothesenwechsel während der Identifikation endlich ist. Die Klasse der Funktionen, die von zuverlässigen Inferenzmethoden identifiziert werden können, hat gegenüber EX den Vorteil, daß sie unter Vereinigung abgeschlossen ist, während es bei EX passieren kann, daß zwei Funktionsmengen identifizierbar sind, ihre Vereinigung aber nicht identifizierbar ist.

V.4. Popper'sche Einschränkung

Eine Beschränkung, die auch als poppersch bekannt ist, wurde 1979 von Case und Ngo-Manguelle eingeführt. Sie verlangt, daß alle Hypothesen, die von einer Identifikationsmethode ausgewählt werden, totale, rekursive Funktionen sein müssen. Es ist also nicht erlaubt, daß eine Funktion, die als Hypothese ausgewählt wird, an irgendeiner Stelle nicht definiert ist. Der Name dieser Einschränkung ist auf den deutschen Philosophen Karl Popper zurückzuführen, der der Ansicht war, daß wissenschaftliche Hypothesen widerlegbar sein müssen, was nur dann gewährleistet ist, wenn die Hypothese eine totale Funktion ist, da der Wert "undefiniert", der bei einer partiellen Funktion auftreten kann weder wahr noch falsch und somit nicht widerlegbar ist. Die totalen, rekursiven Funktionen, die von Popper'schen Maschinen erkannt werden, sind genau die NV-vorhersagbaren Funktionen.

V.5. Konservativität

Eine Identifikationsmethode wird als konservativ bezeichnet, wenn sie nur dann die Hypothese wechselt, wenn die bisherige Hypothese mit den gelesenen Beispielen nicht übereinstimmt. So lange die aktuelle Hypothese zu den gelesenen Beispielen passt, wird sie beibehalten.
Identifikation durch Aufzählung kann als konservative Identifikationsmethode implementiert werden.

V.6. klassenerhaltende Identifikationsmethoden

Eine Identifikationsmethode M heißt klassenerhaltend genau dann, wenn für jede Beispielsequenz für eine Funktion aus einer Klasse von totalen, rekursiven Funktionen gilt, daß jede Hypothese, die von der Inferenzmethode während des Inferenzvorgangs angenommen wird, derselben Klasse angehört.
Ein Beispiel für klassenerhaltende Identifikation ist die Identifikation durch Aufzählung. Diese Identifikationsmethode kann nur Hypothesen aus einer vorher berechneten geordneten Liste ausgeben und ist somit klassenerhaltend bezüglich der Klasse der in dieser Liste gespeicherten Funktionen.

V. 7. Beschränkung auf möglichst gute Hypothesen

Häufig ist es sinnvoll, von einer Identifikationsmethode zu verlangen, daß sie die nach einem festgelegten Kriterium bestmögliche Hypothese ausgeben soll. Hier kommen unter anderem die oben beschriebenen Kriterien für Einfachheit und Genauigkeit zusammen mit der Methode der Identifikation durch Aufzählung zum Einsatz.

VI. Universelle Inferenzmethoden

Besonders interessant ist es natürlich, Identifikatonsmethoden zu finden, die universell einsetzbar sind, und mit kleinen Modifikationen für beliebige Regel- und Hypothesenklassen verwendet werden können. Hierzu bietet sich beispielsweise die Identifikation durch Aufzählung an, die an unterschiedliche Regelklassen angepasst werden kann, indem ihr verschiedene Hypothesenlisten, also Aufzählungen von totalen, rekursiven Funktionen oder rekursiven Sprachen übergeben werden. Es gibt aber auch eine Reihe anderer Inferenzmethoden, die unversell einsetzbar sind. Damit diese Methoden irgendwann zu einer Entscheidung kommen, muß eine Schranke für den Verbrauch irgendeiner Ressource gesetzt werden, die universell, also unabhängig von den zu testenden Hypothesen ist.
Zwei verschiedene Methoden, die auf dem Prinzip der Identifikation durch Aufzählung beruhen, eine obere Schranke für den Verbrauch bestimmter Ressourcen setzen, und bei Überschreitung der Schranke die Berechnungen für eine bestimmte Hypothese abbrechen, werden im Folgenden vorgestellt:

VI. 1. h-einfache Funktionen

Eine Ressource, die beschränkt werden kann, ist die Anzahl der Rechenschritte. Hierzu wird eine totale, rekursive Funktion h(x) verwendet. Eine Funktion f wird genau dann h-einfach genannt, wenn es ein Programm p_i gibt, das f berechnet, so daß p_i für alle bis auf endlich viele Eingaben x maximal h(x) Schritte läuft.
h-einfache Funktionen können leicht im Limes identifiziert werden, indem für eine Eingabe y_1, y_2, ..., y_n alle Paare <i, k> in einer festen Reihenfolge durchsucht werden, bis das erste gefunden wird, für das gilt $p_i(x) = y_x$
für alle x □ n, und p läuft maximal max[k, h(x)]-Schritte auf der Eingabe x für x □ n + 1. Wenn das erste solche Paar <i, k> gefunden wurde, so wird die Hypothese $p_i(n+1)$ ausgegeben.
Diese Inferenzmethode bricht die Berechnungen für eine Hypothese ab, wenn sie nicht in h(x) Schritten zum Ziel gelangt. Wenn k größer wird, als h(x) können die Berechnungen natürlich wieder aufgenommen werden.
Funktionsmengen aus der Klasse NV sind Untermengen der h-einfachen Funktionsmengen, wenn h sich über alle rekursiven Funktionen in einer Variable erstreckt.

VI.2. h-vertrauenswürdige Funktionen

h-vertrauenswürdige Funktionen unterscheiden sich von den h-einfachen Funktionen dadurch, daß die Schranke h von zwei Variablen x und y abhängt. h(x,y) wird a-priori-Grenze genannt, und eine Funktion f heißt genau dann h-vertrauenswürdig, genau dann, wenn ein Programm p_i existiert, das f berechnet, so daß für alle bis auf endlich viele Eingaben x gilt, wenn y = f(x), dann läuft p_i für genau h(x,y) Schritte auf der Eingabe.
Alle Mengen von Funktionen, die von zuverlässigen Identifikationsmethoden erkannt werden können, sind Untermengen von h-vertrauenswürdigen Funktiosmengen, wenn h sich über alle rekursiven Funktionen in zwei Variablen erstreckt.

VI.3. Methoden, die an verschiedene Regelklassen angepasst werden können

Von Jantke wurde 1979 das Prinzip der sogenannten strategischen Operatoren vorgestellt. Strategische Operatoren sind Inferenzmethoden, die nach dem Prinzip der Identifikation durch Aufzählung funktionieren, ein Programm p übergeben bekommen, das eine Reihenfolge für eine bestimmte Klasse U von Funktionen berechnen kann, und damit in der Lage sind, eine geordnete Funktionsliste zu erzeugen und das Prinzip der Identifikation durch Aufzählung auf beliebige Hypothesenklassen anzuwenden.
Jantke unterscheidet selektive und konstruktive strategische Operatoren. Selektive strategische Operatoren sind solche, die nur Elemente aus der Klasse U, die auch vom Programm p aufgezählt wurden, als Hypothesen verwenden. Ein konstruktiver, strategischer Operator konstruiert Hypothesen, indem er, abhängig von der Eingabe rekursiv Elemente aus U auswählt.

Jantke konnte nachweisen, daß konstruktive, strategische Operatoren in einigen Fällen stärker sein können, als selektive. Weiterhin sind konstruktive strategische Operatoren für manche Aufzählungen exponentiell effizienter in der Anzahl von Hypothesenwechseln als selektive.

VII. Komplexität von Inferenzmethoden

Ein axiomatischer Ansatz, um ein Komplexitätsmaß für Inferenzmethoden einzuführen stammt von Daley und Smith. Um ein Komplexitätsmaß für eine Menge von Inferenzprozeduren, aufzustellen, wird jeder Prozedur aus der Menge ein Funktional mit den folgenden Eigenschaften zugeordnet:

- Die Komplexität einer Inferenzprozedur ist genau dann definiert, wenn die Prozedur konvergiert
- Das Komplexitätsfunktional ist exakt auf den Beispielsequenzen definiert, auf denen auch die Inferenzprozedur definiert ist.
- Es gibt eine, rekursive Prozedur, mit deren Hilfe bestimmt werden kann, ob eine gegebene Inferenzprozedur auf einer bestimmten Eingabe eine bestimmte Komplexität besitzt.
- Das Komplexitätsfunktional benötigt genau die gleichen Eingaben um zu konvergieren, wie die Inferenzprozedur
- Die Anzahl der Hypothesenänderungen ist eine absolute, untere Grenze für die Komplexität einer Inferenzmethode.

Daley und Smith konnten mit Hilfe dieser Definition nachweisen, daß es Funktionsmengen gibt, die beliebig schwer zu inferieren sind, da es für sie keine effiziente Inferenzprozedur gibt.

VIII. Induktive Inferenz in der Praxis

Die universellen Inferenzmethoden, die in den letzten Abschnitten beschrieben wurden, sind für den praktischen Einsatz so noch nicht geeignet, da sie eine zu hohe Komplexität besitzen. Um das Prinzip der induktiven Inferenz in der Praxis erfolgreich anzuwenden sind einige Änderungen notwendig.

VIII.1. Suchräume

Statt bei der Identifikation durch Aufzählung die Hypothesen in einer einfachen Liste nach einer festgelegten Reihenfolge anzuordnen, bietet es sich häufig an, komplexere Datenstrukturen, wie zum Beispiel Baumstrukturen zu verwenden, die effizienter durchsucht werden können. Das ist insbesondere dann von Vorteil, wenn ein Problem so aufgebaut ist, daß, immer dann, wenn sich eine Hypothese als falsch herausstellt, mehrere andere Hypothesen, die ein bestimmtes Merkmal aufweisen ebenfalls nicht mehr in Frage kommen. Sind die Hypothesen beispielsweise in einer Baum- oder Graphenstruktur angeordnet, so können in solchen Fällen Brücken, die zwei Zusammenhangskomponenten des Graphen oder Baums verbinden, gelöscht werden, wodurch sich der zu durchsuchende Raum schnell verkleinert.
Ein Beispiel für die erfolgreiche Anwendung von Suchräumen sind Hypothesenklassen, deren Verhältnis untereinander als Spezialisierung oder Generalisierung bezeichnet werden kann. Kann ausgeschlossen werden, daß eine generelle Hypothese zutrifft, so fallen damit automatisch auch alle Unterhypothesen weg.
Angenommen, eine Inferenzmethode soll Klassen von Gegenständen erraten. Eine generelle Hypothese wäre dann beispielsweise "die gesuchte Klasse umfasst alle Arten von Obst" und passende Unterhypothesen "es wird nach der Klasse Apfel gesucht" beziehungsweise "es wird nach der Klasse Birne gesucht". Taucht nun in der positiven Beispielsequenz zum Beispiel die Aussage "in der gesuchten Klasse kommt das Objekt Turnschuh vor" auf, so kann die generelle Hypothese, daß es sich bei allen Objekten um Obst handelt, verworfen werden, und es können natürlich auch die beiden Unterhypothesen nicht mehr zutreffen und dürfen daher ohne irgendwelche zusätzlichen Tests aus dem Suchraum gelöscht werden. Tritt umgekehrt in der Sequenz von Gegenbeispielen ein Objekt auf, für das eine der Unterhypothesen zutrifft, hier also zum Beispiel "in der gesuchten Klasse kommt kein grüner Apfel vor", so lässt sich daraus schließen, daß dann auch die generelle Hypothese "die gesuchte Klasse umfasst alle Arten von Obst" nicht zutreffen kann, da die Klasse der gesuchten Gegenstände sonst natürlich auch alle Arten von Äpfeln und Birnen beinhalten müßte.
Dieses Verfahren wird im Englischen als "concept learning" bezeichnet und lässt sich wie folgt mathematisch beschreiben: Wenn A und B Hypothesen sind, so ist A Generalisierung von B, genau dann, wenn es eine Substitution □ gibt, die auf die Beispiele für A angewendet werden kann, so daß □(A) eine Untermenge der Beispiele für B ist. Widerspricht nun ein positives Beispiel der

Generalisierung A, so muß es auch der Unterhypothese B widersprechen und beide Hypothesen können verworfen werden. Stimmt ein negatives Beispiel hingegen mit der Unterhypothese B überein, so können sowohl B als auch A verworfen werden, da das Negativbeispiel dann ebenfalls mit A übereinstimmen muß, wodurch A ungültig wird.

Eine Erweiterung dieses Konzept ist der von Mitchel entwickelte Algorithmus "version spaces". Er beruht darauf, immer die generellsten Hypothesen abzuspeichern, die auf keins der Gegenbeispiele zutreffen, sowie die speziellsten Hypothesen, die noch alle Positivbeispiele beinhalten. Je länger die Listen von Positiv- und Negativbeispielen werden, desto mehr Hypothesen werden aus der Liste gestrichen, bis nur noch eine übrigbleibt. Ist dieser Zustand erreicht, so ist eine Hypothese eindeutig identifiziert.

VIII.2. Suche nach verbotenen Features

Eine noch komplexere Ordnung der Hypothesenmenge wird verwendet, um nach sogenannten "verbotenen Features" zu suchen, die eine Sequenz von Gegenbeispielen gemeinsam hat. Ein verbotenes Feature ist normalerweise ein bestimmter Bestandteil einer Hypothese, der dazu führt, daß alle Hypothesen, in denen er enthalten ist, zwangsläufig falsch sein müssen. Solche Hypothesen können dann ausgeklammert werden, um die Suche effizienter zu machen.
Wharton beschreibt eine Anwendung dieses Prinzips auf kontextfreie Grammatiken. Er wendet eine Variation des Backtracking-Prinzips an, sortiert kontextfreie Grammatiken aufsteigend nach Komplexität und testet sie dann mit den gegebenen Beispielen. Sobald eine Inkompatibilität auftritt, wird eine Diagnoseroutine aufgerufen, die die erste Produktion in der Grammatik sucht, die geändert werden muß, um die Inkompatibilität zu verhindern. Als nächstes springt der Algorithmus zu der ersten Grammatik, die die fragliche Produktion nicht enthält und fährt mit seiner Arbeit fort. Auf diese Art und Weise wird der Suchraum deutlich eingeschränkt was die Effizienz der Methode im Vergleich zur Identifikation durch Aufzählung erhöht.

VIII.3. Hill Climbing

Eine weitere Methode zur Einschränkung des Suchraums wird als "Hill Climbing" bezeichnet. Beim "Hill Climbing" wird eine Hypothese gesucht, die besser ist, als alle Hypothesen in ihrer unmittelbaren Umgebung der Größe n in einer geordneten Liste und diese nur dann gewechselt, wenn beim Vergrößern der Umgebung eine bessere Hypothese gefunden wurde. Um die Güte einer Hypothese festzulegen, wird in der Regel eine Bewertungsfunktion benutzt. Ein Einsatzgebiet dieser Methode sind insbesondere die stochastischen Sprachen, also solche, bei deren Grammatiken für jede Produktion eine gewisse Wahrscheinlichkeit angegeben wird, mit der diese Produktion beim Bilden eines Wortes zum Einsatz kommt.

VIII.4. Erraten von Lösungen

Die bisher vorgestellten Methoden haben gemeinsam, daß sie in keinem Fall Hypothesen überspringen, die möglicherweise korrekt sein könnten. Dadurch arbeiten sie zwar sehr zuverlässig aber auch entsprechend langsam, da die Größe des Suchraums exponentiell mit der Länge der zu suchenden Hypothese wächst.
Es gibt aber eine Klasse von Methoden, die auf einem anderen Prinzip basieren. Diese werden "konstruktiv" genannt. Statt nach Hypothesen zu suchen, die auf keinen Fall korrekt sein können, suchen sie nach Features, die wahrscheinlich in einer korrekten Hypothese enthalten sind. Diese Features werden dann zu einer Hypothese zusammengesetzt. Auf diese Art und Weise wird die Anzahl der zu durchsuchenden Hypothesen stark reduziert, so daß in günstigen Fällen sogar nur eine einzige übrig bleibt.

Methoden, die sich in diese Klasse einordnen lassen, können grob in heuristische und nicht heuristische Methoden unterteilt werden. Nicht heuristische Methoden zeichnen sich dadurch aus, daß ihre möglichen Ergebnisse leicht zu charakterisieren sind. Im Gegensatz dazu lässt sich bei heuristischen Methoden häufig nur eine sehr vage Aussage darüber treffen, welche Art von Ergebnissen sie liefern. Hier jeweils ein Beispiel für beide Arten von Methoden:

Eine heuristische Methode wurde 1972 von Biermann und Feldman vorgestellt. Mit der k-Tails-Methode können nichtdeterministische, endliche Automaten, die aus einer Eingabe lesen und in eine Ausgabe schreiben, anhand ihrer Ein- und Ausgabedaten erzeugt werden. Hierzu muß vom Benutzer eine natürliche Zahl k angegeben werden, die bestimmt, wann zwei Zustände des Automaten als

gleich angesehen werden sollen. Zwei Zustände gelten genau dann als gleich, wenn sie sich auf allen Wörtern der Länge □ k gleich verhalten. Der Algorithmus konstruiert einen generellen endlichen Automaten zu der gegebenen Beispielsequenz und teilt die Zustände des Automaten danach in Blöcke ein. Zwei Zustände sind genau dann in demselben Block, wenn sie mit Wörtern der Länge □ k aus der Beispielsequenz nicht zu unterscheiden sind. Die Zustände innerhalb eines Blocks werden als nächstes zu einem einzigen Zustand zusammengefasst. Wählt man jetzt k ausreichend groß und hat außerdem eine ausreichend lange Beispielsequenz, so wird ein korrekter deterministischer Automat erzeugt.

Eine nicht heuristische Inferenzmethode wurde 1972 von Dana Angluin entwickelt und beschäftigt sich mit sogenannten "pattern"-Sprachen. "Pattern" lässt sich mit "Muster" übersetzen und ein Pattern besteht aus einer Aneinanderreihung von beliebig vielen Konstanten und Variablen. Korrekte Patterns wären beispielsweise 242xx22 oder 452z7zy56y. Eine Sprache, die aus einem Pattern erzeugt wird, umfasst alle möglichen Ersetzungen der im Pattern auftretenden Variablen. Hierbei muß die gleiche Variable immer durch die gleiche Konstante beziehungsweise die gleiche Sequenz von Konstanten ersetzt werden. Wörter der von 2x2xyy generierten Sprache wären also zum Beispiel 222266 oder 2342349797, nicht aber 242566, da im letzten Fall die Variable x einmal durch die Konstante "4" und einmal durch die Konstante "5" ersetzt wurde. Dana Angluin stellte 1980 einen Algorithmus vor, der eine optimale Pattern-Sprache zu einer gegebenen Beispielsequenz inferiert. Im Spezialfall von Sprachen, die nur eine Variable beinhalten, konnte bewiesen werden, daß der Algorithmus in Polynomialzeit abgearbeitet werden kann. Weitere Spezialfälle von Pattern-Sprachen, die mit Polynomzeit-Algorithmen inferiert werden können, sind die erweiterten regulären Pattern-Sprachen, bei denen keine Variable in einem Pattern mehrmals vorkommt und jede Variable nicht nur durch Konstanten sondern auch durch das leere Wort ersetzt werden kann, und die non-cross-Pattern-Sprachen, bei denen alle Vorkommen einer Variable entweder rechts oder links von allen Vorkommen aller anderen Variablen stehen, Patterns wie xyxy oder xxyx also nicht zulässig sind, sondern nur solche wie xxyzz. Algorithmen, die diese Sprachen inferieren, beginnen mit einem universellen Pattern und verfeinern es dann anhand der gegebenen Beispielsequenz. Hierdurch kommt es zu einer korrekten Identifikation im Limes sowohl für die erweiterten regulären Pattern-Sprachen als auch für die non-cross-Pattern-Sprachen

IX. Anwendungen

In der Praxis gibt es bereits eine Reihe von Bereichen, in denen induktive Inferenz erfolgreich eingesetzt wird. Feigenbaum und Buchanan benutzten erfolgreich Inferenzmethoden, um damit Regeln für die Stabilität bestimmter chemischer Moleküle aus massenspektrographischen Beispieldaten zu gewinnen.
Ein anderer interessanter Ansatz stammt von Syklossy und Sykes, die Induktive Inferenz dazu anwenden, einen Roboter Lösungen, die er für spezielle Probleme gefunden hat, generalisieren zu lassen.
Weitere Anwendungen betreffen das Lösen von Rätseln anhand von Beispielen oder das erzeugen von Programmen aus vom User angegebenen Beispieldaten.

X. Wie geht es weiter?

In dieser Arbeit sowie in dem zugrundeliegenden Beitrag von Angluin und Smith fehlen definitive Antworten zu ganz grundlegenden Fragen im Bereich der induktiven Inferenz, wie zum Beispiel, was die beste Inferenzmethode zu einem bestimmten Problem ist, welche Maßstäbe für das Verhalten einer Inferenzmethode wirklich relevant sind, und welche Fähigkeiten einer Inferenzmethode gegebenenfalls mit welchen Zugeständnissen im Bezug auf Speicherplatz, Rechenzeit, Hintergrundwissen, Genauigkeit, Datenmenge etc. erkauft werden müssen.
Die wichtigste noch offene Frage betrifft den Zusammenhang zwischen den hier gezeigten abstrakten Lösungen und den konkreten Anwendungen in der Praxis. Bisher ist nicht sichergestellt, daß sich zu der immer größer werdenden Anzahl theoretischer Resultate im Bereich der induktiven Inferenz auch tatsächlich praktische Anwendungen finden lassen.
Es stellt sich die Frage, wo die Grenzen der Induktive Inferenz nach dem Gold'schen Paradigma der Identifikation im Limes liegen, da beim menschlichen Lernen und Schlußfolgern sowohl induktive als auch deduktive Prozesse beteiligt sind, und auch das Induktive Lernen nicht ausschließlich auf das Lernen im Limes beschränkt ist. Gerade in Bereichen, wo Informationen in einer Vielzahl unterschiedlicher Formen vorliegen, oder in denen viele Faktoren untereinander abhängig sind, ist die induktive Inferenz nach dem Paradigma von Gold nicht die Methode der Wahl.

Literatur:
[1] Angluin, D. and Smith, C. (1983). Inductive inference: Theory and methods. Computing Surveys, 15(3):237--269.