Lernalgorithmen in der Zeitreihenanalyse, eine empirische Untersuchung

Paul Fischer[1] und **Klaus-Uwe Höffgen**[2]
Lehrstuhl Informatik II, Universität Dortmund
D-44221 Dortmund, Germany
paulf/hoeffgen@goedel.informatik.uni-dortmund.de

Zusammenfassung

Das Ziel dieses Projekts war es, anhand von empirischen Untersuchungen klassische statistische Verfahren und aktuelle Methoden des Maschinellen Lernens mit einem Ansatz zu vergleichen, der in der Arbeitsgruppe entworfen und theoretisch analysiert wurde. Implementiert wurden fünf Verfahren, einige davon in verschiedenen Varianten: Feed-Forward Neuronale Netze, Entscheidungsbäume, Bayes Entscheidungen, die auf Chow-Expansionen beruhen, Harmonische Analyse und die Methode des Nächsten Nachbarn. Als Referenzmaßstab wurden Vorhersagen herangezogen, die den Trend oder den Mittelwert der letzten letzten Beobachtungen vorhersagten. Als Daten standen 16 Zeitreihen von Aktien- und Devisenkursen zur Verfügung. Jede der Zeitreihen bestand aus 2000 Daten, von denen die ersten 1500 zum Training und die restlichen 500 für den Vergleich der Verfahren dienten. Dabei zeigte es sich, daß die naiven Referenzverfahren einen recht guten Prüfstein darstellten. Die Bayes-Entscheidungen und die Entscheidungsbäume erwiesen sich als besonders stark und übertrafen die Referenzmethoden fast immer. Neuronale Netze und die Methode des nächsten Nachbarn waren etwa genausogut, während die Harmonische Analyse für kurzfristige Vorhersagen schlechter und für langfristige besser war. Bei Entscheidungsbäumen und Neuronalen Netzen fiel auf, daß kleine Bäume bzw. Netze bessere Ergebnisse lieferten als große.

Den Anlaß für diese Untersuchung bildeten theoretische Analysen von Lernverfahren für probabilistische Konzepte, die von Anoulova, Fischer, Pölt und Simon, [3, 1] durchgeführt wurden. Die Arbeit wurde zum Teil im Rahmen einer Projektgruppe an der Universität Dortmund durchgeführt, [2]. Im Vordergrund standen dabei Bayes-Entscheidungen, die aufgrund von parametrisierten Verteilungen getroffen wurden. Für einige Klassen von Verteilungen, Chow-Expansionen und Bahardur-Lazarsfeld-Expansionen, wurden obere Schranken

[1]Unterstützt durch die Deutsche Forschungsgemeinschaft, Beihilfe We 1066/6-1.
[2]Der Autor dankt für die Unterstüzung durch das Bundesministerium für Forschung und Technologie, Beihilfe 01IN102C/2.

für die Stichprobengröße in Abhängigkeit von der gewünschten Vorhersagegenauigkeit bestimmt. Erste empirische Tests, zum Beispiel im Bereich der Ziffernerkennung, [4], zeigten, daß diese Schranken wirklich eine Worst-Case Analyse darstellen, und man in der Praxis mit wesentlich kleineren Stichproben bereits sehr gute Ergebnisse erzielt. Unser Ziel war nun, die Güte dieser Verfahren an klassischen Methoden der Statistik und des Maschinellen Lernens zu messen. Jedes der Verfahren konnte anhand von Parametern individuell konfiguriert werden. Wir geben nun eine Übersicht über die Verfahren und die zugehörigen Parameter.

Chow-Expansionen: Dies sind Wahrscheinlichkeitsverteilungen, die eine Verallgemeinerung der Markov-Eigenschaft zulassen. Sie lassen sich als Produkt von bedingten Verteilungen schreiben, wobei jeder Faktor nur Abhängigkeiten von bis zu k vorangehenden Variablen zuläßt.

$$P(x_1, \ldots, x_n) = P(x_1, \ldots, x_k) \cdot P(x_{k+1}|x_k, \ldots, x_1) \cdot \cdots \cdot P(x_n|x_{n-1}, \ldots, x_{n-k})$$

Für jeden möglichen Vorhersagewert wird eine solche Chow-Expansion aufgestellt. Während der Trainingsphase werden Schätzwerte für die Faktoren der Verteilungen berechnet. Vorhergesagt wird derjenige Wert, dessen geschätzte Verteilung die größte Wahrscheinlichkeit berechnet. Als Parameter kommen hier k und die Mengen der Bedingungsvariablen in Frage.

Nächster Nachbar: Hierbei werden jeweils k aufeinanderfolgende Trainingsdaten zu Tupeln zusammengefaßt. Bei der Vorhersage wird aus den letzten Beobachtungen ebenfalls ein k Tupel gebildet und das dazu nächstgelegene Tupel aus Trainingsdaten gesucht. Die Vorhersage ist dann die Klassifikation, die zum Trainingstupel gehört. Eine Modifikation besteht darin die l nächsten Nachbarn zu bestimmen und den Wert vorauszusagen, der bei ihnen an häufigsten auftritt. Die Parameter sind hier k und l.

Neuronale Netze: Wir haben Feed-Forward-Netze betrachtet, die mit Backpropagation trainiert wurden. Als Eingabe dienten die beobachteten Daten. Es gab mehrere Ausgabeknoten, einen für jeden möglichen Vorhersagewert. Vorhergesagt wurde derjenige, dessen Knoten die höchste Aktivierung hatte. Als Parameter wurden hier die Anzahl der Schichten und die Anzahl der Knoten pro Schicht gewählt. Außerdem wurden verschiedene Lernregeln angewandt.

Entscheidungsbäume: Implementiert wurden ein einfaches Verfahren zum Aufbau von Entscheidungsbäumen und ID3 (siehe [5]). Um die Komplexität der Bäume zu beschränken, wurde die bekannte "Fenstertechnik" angewendet: es wird aus der Trainingsmenge ein Teilmenge (Fenster) ausgewählt und zum Baumaufbau herangezogen. Der resultierende Baum wird dann auf der übrigen Trainingsmenge getestet. Bei zu großem Fehler (einstellbarer Parameter) dieses Baums wurde die Teilmenge erweitert und das Verfahren wiederholt.

Harmonische Analyse: Auf der Zeitreihe wurde eine wurde eine Harmonische Analyse durchgeführt. Von den berechneten Spektralkoeffizienten wurden einige ausgewählt und die von ihnen definierte trignometrische Funktion wurde zur Vorhersage benutzt. Variiert wurde die Auswahl der Koeffizienten.

Naive Methoden: Diese dienten als Referenzverfahren. Sie beruhen darauf, bei einer Zeitreihe den zuletzt beobachteten Wert oder den Durchschnitt der zuletzt beobachteten Werte vorherzusagen. Parameter waren hier die Anzahl der zur Durchschnittsbildung herangezgenen Daten.

Für den Vergleich entschieden wir uns für ein eng begrenztes Gebiet, nämlich die Vorhersage von univariaten Zeitreihen. Uns standen insgesamt 16 Zeitreihen aus mit Aktien- und Devisenkursen zur Verfügung, jeweils die täglichen Mittelkurse. Die Zeitreihen deckten einen Zeitraum von zehn bis zwanzig Jahren ab.

Unser Ziel war die Voraussage des Trends für den jeweils nächsten Tag. Es gab fünf Trendwerte: *stark fallend, fallend, wenig verändert, steigend und stark steigend.* Für jede Zeitreihe erfolgte die Einteilung so, daß jede der Klassen (soweit möglich) gleich groß war. Von den Zeitreihen wurden die letzten zwei Jahre abgetrennt, also etwa 500 Daten. Auf den Anfangsstücken wurden die Lernverfahren dann trainiert. Jedes der Verfahren wurde dabei zunächst für jede der Zeitreihen individuell optimiert, d.h es wurden besonders gute Parametersätze gesucht.

Anschließend wurden die Verfahren ohne weiteres Training auf den Testdaten verglichen. Die Verfahren erhielten sukzessive die unbekannten Daten und mußten den Trend für den jeweils nächsten Tag vorhersagen. Zur Gütebewertung wurden mehrere Kriterien herangezogen.

(F) Die Anzahl der Fehler, d.h. die Anzahl der Vorhersagen, die nicht den korrekten Trendwert trafen.

(GF) Eine gewichtete Fehleranzahl, d.h. eine Vorhersage, die nur um eine Klasse abweicht (z.B. steigend statt stark steigend), wurde geringer gewichtet als eine, die sich um mehrere Klassen irrt.

(T) Anzahl der Fehler im Trend, d.h. Vorhersage (stark) steigend, obwohl in Wirklichkeit (stark) fallend zutraf.

Es ist bekannt, daß Devisen- und besonders Aktienkurse sehr schwer vorherzusagen sind. Diese Daten weisen kaum Struktur auf. Man sollte daher bei einer Klassifizierung mit fünf Werten nicht wesentlich mehr als 20% Genauigkeit beim Kriterium (F) erwarten.

Vergleich mit den Naiven Methoden: Als stärkste Naive Methode erwies sich diejenige, die den Trendwert voraussagte, der in den letzten Beobachtungen am häufigsten vorkam, je nach Zeitreihe waren das 10 bis 40

Beobachtungen. Die Chow-Expansion und die Entscheidungsbäume erwiesen sich als die stärksten Verfahren. Sie wurden nur bei zwei Zeitreihen von den naiven Methoden geschlagen und waren ansonsten um 2 bis 8 Prozentpunkte besser als die beste naive, durchschnittlich um 4 Prozentpunkte. Dies galt für alle Gütekriterien (F), (GF) und (T). Die Neuronalen Netze und die Methode des nächsten Nachbarn zeigten sich etwa so stark wie Naiven Methoden. Die Methode des nächsten Nachbarn wies dabei wesentlich größer Schwankungen auf als die Neuronalen Netze. Die Harmonische Analyse zeige sich schwächer als die naiven Methoden. Allerdings holte sie bei der Langfristvorhersage deutlich auf.

Entwicklung der Vorhersagegüte: Alle Methoden zeigten bei den ersten 20 bis 50 Vorhersagen ein deutlich besseres Verhalten als bei den weiteren. Sowohl für Chow-Expansionen, als auch für Entscheidungsbäume gilt aber, daß sie bei bis zu 500 Vorhersagen besser blieben als die naiven Methoden. Um diesem Nachlassen der Genauigkeit entgegenzuwirken, wurden weitere Testläufe durchgeführt, in denen nach jeweils 20 Vorhersagen eine (kurze) Trainigsphase stattfand, ohne allerdings die Parameter wesentlich zu ändern (bei den Netzen wurde beispielsweise die Topologie nicht verändert, sondern nur die Gewichte). Dies erhöhte zwar die Genauigkeit, aber die Genauigkeit der ersten 20 bis 50 Vorhersagen konnte auch so nicht bis zum Ende durchgehalten werden. Wir vermuten, daß, um die Anfangsgenauigkeit aufrechtzuhalten, immer wieder komplette Trainingsphasen nötig sind, die eventuell auch neue Parametersätze erzeugen.

Bemerkungen zu den Methoden: Das gute Abschneiden der Chow-Expansion war nach den Erfahrungen, die wir mit dieser Methode bei der Ziffernerkennung gesammelt hatten, nicht überraschend. Überraschend war aber, daß Entscheidungsbäume deutlich besser arbeiteten als Neuronale Netze. Bei den beiden letztgenannten Methoden fiel außerdem auf, daß kleine Bäume (mit 2 bis 4 Attributen) und kleine Netze (eine Hidden-Schicht mit 1 bis 2 Neuronen) bessere Ergebnisse lieferten als große. Offensichtlich liefern, zumindest für die untersuchten Zeitreihen, einfache, aber nicht zu naive Methoden recht gute Ergebnisse. Versuche mit größeren Netzen und Bäumen führten zum Teil zu besseren Ergebnissen auf den Trainingsdaten. Auf den Testdaten allerdings waren die Ergebnisse schlechter. Das deutet darauf hin, daß diese Netze bereits ,,auswendiglernen" und nicht mehr so gut generalisieren.

Weitere Tests: Der Versuch die Methoden nicht für jede Zeitreihe individuell zu optimieren, sondern das Training auf einigen oder allen Reihen durchzuführen, brachte kein Verbesserung. Eine Ausweitung des Vorhersagehorizonts von einem Tag auf 125 Tage brachte einige Änderungen. Die Chow-Expansion wurde geringfügig schlechter, blieb aber besser als die naiven Methoden. Entscheidungsbäume, Neuronale Netze und der Nächste Nachbar wurden

geringfügig besser. Lediglich die Harmonische Analyse wurde deutlich besser und übertraf sogar die naiven Methoden. Bei einigen Zeitreihen war sie sogar die beste Methode.

Danksagung: Wir bedanken uns für die Implementierung und die Durchführung der Testläufe bei den Mitgliedern der Projektgruppe DELPHI: Mahsuni Basci, Julia Brasse, Elif Dilek, Arnfried Griesert, Michael Hanhörster, Alexander Holland, Frank Jagla, Sevim Kosan, Klaus Luttmann, Tobias Padberg, Michael Reichelt, Lutz Schneider.

Außerden danken wir der Westdeutschen Landesbank für ihre Unterstützung und die Bereitstellung der Daten.

Literatur

[1] S. Anoulova, P. Fischer, S. Pölt, and H. U. Simon. PAB-decisions for Boolean and real-valued features. In *Proc. 5th Annu. Workshop on Comput. Learning Theory*, pages 353–362. ACM Press, New York, NY, 1992.

[2] P. Fischer and K.-U. Höffgen, editors. *Endbericht der Projektgruppe DELPHI.* Fachbereich Informatik, Universität Dortmund, 1994.

[3] P. Fischer, S. Pölt, and H. U. Simon. Probably almost Bayes decisions. In *Proc. 4th Annu. Workshop on Comput. Learning Theory*, pages 88–94. Morgan Kaufmann, San Mateo, CA, 1991.

[4] S. Pölt and H.-U. Simon, editors. *Endbericht der Projektgruppe Lamex-Z.* Fachbereich Informatik, Universität Dortmund, 1994.

[5] J. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.