Stephan Wöbbeking

Design und Implementation eines Softwaresystems für die Klassifikation und Prognose von Zeitreihen

Diplomarbeit

Examicus

Wöbbeking, Stephan:
Design und Implementation eines Softwaresystems für die Klassifikation und Prognose von Zeitreihen - 2001. - 148 S. Mittweida, Hochschule Mittweida (FH) - University of Applied Sciences, Fachbereich MPI, Diplomarbeit, 2001

Referat:
Das Ziel der Diplomarbeit ist es, ein Vorhersagesystem für Zeitreihen zur Verfügung zu stellen. Es ist speziell ausgelegt auf die Bedürfnisse von Handel und Wirtschaft und berücksichtigt die Besonderheiten in den vorgelegten Datenbeständen. Das System untersucht die verschiedenen Merkmale in den Reihen und zeigt einige Lösungsansätze auf Basis einer Klassifikation auf. Diese Vorgehensweisen müssen in der Praxis getestet und weiterentwickelt werden. Die vorgelegte Implementierung zeigt einen Weg auf, wie mit dem System S-PLUS strukturiert programmiert werden kann. Es zeigt ebenfalls einen Weg für ein durchgängiges Datenkonzept auf. Die Arbeit umfaßt die drei wesentlichen Komponenten: ein Vorhersagesystem (Prognose), die Kennwertberechnung sowie die Visualisierung.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Quellenausschnitte

Listings

Kapitel 1

Einführung

1.1 Aufgabe

Im Rahmen dieser Diplomarbeit wird ein Problem der Klassifikation und Prognose von Zeitreihen bearbeitet. Es behandelt die Vorhersage von Verkaufszahlen. Ein führender amerikanischer Hersteller von Glühbirnen möchte Produktion und Vertrieb effizienter gestalten. Daher ist es von entscheidender Bedeutung, gute Prognosen über die eintretenden Verkaufsmengen erstellen zu können. Eine breite Auswahl unterschiedlicher Produkte führt zu einer großen Menge auftretender Daten. Pro Monat werden für jeden Artikel in verschiedenen Regionalbereichen die Bestellmengen aufsummiert und in Form von Zeitreihen in einer Datenbank abgelegt. Diese Daten dienen als Ausgangspunkt für die vorgenommenen Untersuchungen. In einer ersten Abstraktionsstufe werden die Verkaufszahlen von dem eigentlichen Produkt gelöst. Es handelt sich nun nur noch um einfache Zeitreihen, deren Verhalten untersucht, klassifiziert und prognostiziert werden soll. Diese Abstraktion wird bereits seitens des Herstellers vorgenommen und soll daher auch nicht im einzelnen erläutert werden. Die folgenden Untersuchungen beziehen sich daher auf Zeitreihen mit den verschiedenen Besonderheiten.

Es werden Wege und Ansätze aufgezeigt, die eher zu einer zufriedenstellenden Vorhersage der Zeitreihen führen können, als dies mit untersuchten Komplettlösungen möglich ist. So stand am Anfang die Beurteilung zweier verfügbarer Systeme, die laut Herstellerangaben bereits gute Ergebnisse liefern können. Nachdem sich jedoch herausgestellt hatte, daß diese Lösungen zu ungenaue Ergebnisse liefern, fiel die Entscheidung, eine eigene Programmierung vorzunehmen. Als Sprache stand hier das in Abschnitt 1.4 charakterisierte Paket S-PLUS zur Verfügung. Es waren bereits wenige Lösungsansätze erarbeitet worden. Diese ordneten sich allerdings nicht in ein Komplettsystem ein, sondern waren vielmehr unzusammenhängende Teilimplementationen für Ausschnitte aus dem Gesamtproblem. Sie mußten sowohl aufbereitet als auch um weitere Teillösungen erweitert werden. Der vorhandene Datenbestand war ungeordnet in verschiedenen Datensammlungen und Formaten verfügbar. Die Erarbeitung eines möglichst einfachen, jedoch universellen Datenkonzeptes wurde angestrebt. Anschließend wurden einige Programmfragmente analysiert, um die umgesetzten Algorithmen herauszufiltern und zu ordnen.

Die Problemanalyse schloß die Analyse der vorhandenen Ideen und die Überprüfung ihrer Effizienz ein. Nach einer ersten Datenanalyse wurde ein Konzept erarbeitet, um Strukturen und Relationen zu erkennen. Innerhalb dieses Systems ist es möglich, die erforderlichen Berechnungen durchzuführen sowie verschiedene Ergänzungen vorzunehmen. Hierbei wurde größtmöglicher Wert auf die Erweiterbarkeit, Wiederverwendbarkeit und Wartbarkeit gelegt, da weitere Untersuchungen durchgeführt werden sollen und diese zu weiteren Ergänzungen führen könnten. Die weitere Analyse der Daten soll mit Hilfe des erstellten Systems leichter durchgeführt werden können. Soweit möglich, werden sinnvoll erscheinende Änderungen direkt in das Programmsystem übernommen.

Das Ziel dieser Diplomarbeit ist es, mathematische Modelle zu entwickeln und in ein Programmsystem umzusetzen. Dieses wird um einfache Berechnungen ergänzt, die die Modelle unterstützen und ergänzen, jedoch grundsätzlich nicht verändern. Der Schwerpunkt liegt somit in der Programmierung eines Systems, mit dessen Hilfe der Nutzer Prognosen erstellen, visualisieren und später auch deren Genauigkeit darstellen kann.

1.2 Problemstellung

Die gestellte Aufgabe läßt sich in zwei große Teilaufgaben gliedern. In einem ersten Teil sollen Vorhersagen für die gegebenen Zeitreihen gemacht werden. Anschließend ist nach einer geeigneten Möglichkeit zu suchen, dem Nutzer diese Daten zugänglich zu machen.

1.2.1 Prognose

Das weit größere der beiden Teilprobleme ist die Prognose der Zeitreihen. Laut der Aufgabenstellung sollen bei Nichtkenntnis der zukünftigen Werte diese möglichst genau prognostiziert werden. Bezüglich der Rechenzeit oder dem Einsatz professioneller Produkte usw. bestehen keine Einschränkungen.

1.2.2 Visualisierung

In einem zweiten Schritt ist die Bereitstellung einer geeigneten Präsentationsmöglichkeit gefordert. Die prognostizierten Werte sollten nicht nur als Zahlenwerte vorliegen, sondern optisch aufbereitet werden und mit dem vergangenen Verlauf der Zeitreihe zusammen visualisiert werden.

1.3 Rahmenbedingungen: Ein Praktikum bei der Firma Siemens

Die vorgelegte Arbeit wurde abgeleitet aus den Tätigkeiten, die im Rahmen eines Praktikums ausgeführt wurden. Dieses Praktikum wurde bei der Firma **Siemens Corporate Research** in Princeton / New Jersey in den Vereinigten Staaten durchgeführt. Es kam im Rahmen des Austauschprogramms **SSP - Siemens students program** zustande. Die Haupttätigkeiten in den USA waren einige Voruntersuchungen zu dem behandelten Thema, der Datenbankanbindung und der Analyse vorgegebener Lösungen.

1.4 Überblick über die zeitliche Entwicklung

Eine immer höhere Rechenleistung moderner Computer ermöglicht heute komplexe Berechnungen, die vor wenigen Jahren noch nicht durchführbar waren.

Ohne diese Entwicklung würde die Bewältigung einer gestellten Aufgabe so lange dauern, daß das erzielte Ergebnis schon nicht mehr von Bedeutung ist, wenn man denn überhaupt zu einem Ergebnis gekommen wäre. So werden viele Problemstellungen, für die bisher keine ausreichende Rechenleistung zur Verfügung stand, heutzutage neu aufgegriffen. Mit Hilfe dieser hohen Leistung werden Versuche unternommen, um die Berechnungen zu erweitern, zu verbessern oder überhaupt erst in einer angemessenen Zeit durchzuführen.

Auf den Bereich der Klassifikation von Zeitreihen zum Zwecke ihrer Einschätzung, Charakterisierung und deren Prognose trifft dies im besonderen zu. Zu dieser Thematik wird mit der vorgelegten Arbeit ein Beitrag geleistet. Es wird mit so umfangreichen Datenbeständen gearbeitet, daß eine Bearbeitung ohne den intensiven Einsatz der Datenverarbeitungstechnik unmöglich ist. Seit diese Verarbeitung jedoch möglich ist, werden Anstrengungen unternommen, neue Systeme zu schaffen und zu erweitern. Diese können umfangreiche statistische Kalkulationen durchführen. So werden manchmal Informationen gewonnen, deren Berechnung bisher nicht möglich war, die für die Lösung der verschiedensten Probleme jedoch hilfreich sein können.

Eine Reihe kommerzieller Produkte sind mittlerweile auf dem Markt, die darauf spezialisiert sind, die Hersteller- und Händlerbedürfnisse zu befriedigen. Es geht meistens darum, das Kundenverhalten besser einzuschätzen und Aussagen für die Zukunft treffen zu können. Zuverlässige Vorhersagen in diesem Zusammenhang sind von beträchtlichem Interesse, da sie in nahezu allen Branchen zu extremer Kosteneinsparung führen. Es sei insbesondere auf die Produktion, den Transport und die Lagerhaltung hingewiesen. Da jedoch nahezu jeder Datenbestand besondere Merkmale aufweist, liefern die untersuchten Komplettpakete keine zufriedenstellenden Lösungen. Mit ihnen wird der Versuch unternommen, die Probleme zu generalisieren, die besser individuell bearbeitet werden sollten.

Es bietet sich die Möglichkeit, jedes mathematische oder statistische Problem mittels einer allgemeinen Programmiersprache wie Pascal, C, Java etc. innerhalb eines ausführbaren Programmes zu lösen. Hierbei besteht jedoch die Problematik, daß diese Programmiersprachen nicht auf den Einsatz in der Mathematik spezialisiert sind und im Sprachstandard nur über sehr geringe mathematische Sprachelemente verfügen. Die Grundlagen sind implementiert, jedoch müssen alle weiterführenden Berechnungen vom Programmierer umgesetzt werden. Dies erfordert einen extrem hohen Aufwand an Personal und Zeit und verursacht dadurch enorme Kosten. Da gerade die Kosten- und Zeiteinsparungen die Gründe sind, warum solche Probleme gelöst werden sollen, ist dies ein unbefriedigendes Vorgehen.

Mittlerweile gibt es zusätzlich Produkte, die zwischen den Komplettlösungen und den Programmiersprachen einzuordnen sind. Vorwiegend handelt es sich hierbei um Programmiersprachen, denen herkömmliche Elemente fehlen, die dafür jedoch über sehr umfangreiche mathematische Funktionsbibliotheken ver-