

DATA-FORECASTER

Erich Steiner
steiner@softopt.de
SoftOpt
www.softopt.de

7. April 2003

Was ist DATA-FORECASTER?

Was ist DATA-FORECASTER?

- eine Software für Datamining

Was ist DATA-FORECASTER?

- eine Software für Datamining
- ein neuer unveröffentlichter Algorithmus

Was ist DATA-FORECASTER?

- eine Software für Datamining
- ein neuer unveröffentlichter Algorithmus
- leicht zu benutzen, einfache Benutzerschnittstelle

Was ist DATA-FORECASTER?

- eine Software für Datamining
- ein neuer unveröffentlichter Algorithmus
- leicht zu benutzen, einfache Benutzerschnittstelle
- viele Anwendungen in
 - ★ Wirtschaft,
 - ★ Medizin,
 - ★ Investment-banking,

- ★ Customer-relationship-management,
- ★ Versicherungen,
- ★ Aufdeckung von Betrug,
- ★ Pharmazeutik.

Versicherungsbeispiel

Wir haben Daten aus der Vergangenheit

Versicherungsbeispiel

Wir haben Daten aus der Vergangenheit

Klient	Alter	Geschlecht	verheiratet	Bildung	Kostenklasse
1	25	Mann	ja	Hochschule	2
2	45	Frau	nein	Lehre	3
3	32	Frau	ja	Abitur	1
...
100000	21	Mann	nein	Mittlere Reife	5

Versicherungsbeispiel

Wir haben Daten aus der Vergangenheit

Klient	Alter	Geschlecht	verheiratet	Bildung	Kostenklasse
1	25	Mann	ja	Hochschule	2
2	45	Frau	nein	Lehre	3
3	32	Frau	ja	Abitur	1
...
100000	21	Mann	nein	Mittlere Reife	5

Klient kommt zu uns, wir wissen sein Alter, Geschlecht, ob er/sie verheiratet ist, den Bildungsgrad, aber wir wissen (noch) nicht die Kostenklasse.

DATA-FORECASTER liefert eine Vorhersage für die Kostenklasse.

DATA-FORECASTER liefert eine Vorhersage für die Kostenklasse.

	Klasse	1:	1.5 %
	Klasse	2:	77.5 %
Zum Beispiel:	Klasse	3:	9,3 %
	Klasse	4:	6.2 %
	Klasse	5:	5.5 %

Hauptfunktionen von DATA-FORECASTER

Hauptfunktionen von DATA-FORECASTER

- Konstruktion eines Vorhersagemodells.

Hauptfunktionen von DATA-FORECASTER

- Konstruktion eines Vorhersagemodells.
- Vorhersage zu welcher Klasse ein Individuum gehört.

Hauptfunktionen von DATA-FORECASTER

- Konstruktion eines Vorhersagemodells.
- Vorhersage zu welcher Klasse ein Individuum gehört.
- Messung der Qualität eines Vorhersagemodells indem Vorhersagen und tatsächliche Klassenzugehörigkeit von Daten aus der Vergangenheit kreuzverglichen werden.

Hauptfunktionen von DATA-FORECASTER

- Konstruktion eines Vorhersagemodells.
- Vorhersage zu welcher Klasse ein Individuum gehört.
- Messung der Qualität eines Vorhersagemodells indem Vorhersagen und tatsächliche Klassenzugehörigkeit von Daten aus der Vergangenheit kreuzverglichen werden.
- Untersuchung, ob die Daten Korrelationen enthalten.

Hauptfunktionen von DATA-FORECASTER

- Konstruktion eines Vorhersagemodells.
- Vorhersage zu welcher Klasse ein Individuum gehört.
- Messung der Qualität eines Vorhersagemodells indem Vorhersagen und tatsächliche Klassenzugehörigkeit von Daten aus der Vergangenheit kreuzverglichen werden.
- Untersuchung, ob die Daten Korrelationen enthalten.
- Bewertung der Qualität einer anderen Vorhersagemethode, Vergleich mit DATA-FORECASTER möglich.

Warum DATA-FORECASTER benutzen?

Warum DATA-FORECASTER benutzen?

- Datamining kann intern in der Firma durchgeführt werden, nicht notwendig vertrauliche Daten extern bearbeiten zu lassen.

Warum DATA-FORECASTER benutzen?

- Datamining kann intern in der Firma durchgeführt werden, nicht notwendig vertrauliche Daten extern bearbeiten zu lassen.
- Es ist verhältnismäßig billig. Nicht mehr notwendig, langwierige Projekte mit Beratern durchzuführen.

Warum DATA-FORECASTER benutzen?

- Datamining kann intern in der Firma durchgeführt werden, nicht notwendig vertrauliche Daten extern bearbeiten zu lassen.
- Es ist verhältnismäßig billig. Nicht mehr notwendig, langwierige Projekte mit Beratern durchzuführen.
- Es ist schnell. Zeit benötigt für Modellkonstruktion
 - ★ 20 Sekunden für ein 1000 X 15 Problem
 - ★ 37 Minuten für ein 1200 000 X 15 Problem

- Entwickelt aus neuesten Forschungsergebnissen. Der zugrundeliegende Algorithmus ist eine Kombination aus Baumkonstruktion, hybridem genetischem Algorithmus und quadratischer Optimierung.

- Entwickelt aus neuesten Forschungsergebnissen. Der zugrundeliegende Algorithmus ist eine Kombination aus Baumkonstruktion, hybridem genetischem Algorithmus und quadratischer Optimierung.
- Die Vorhersagen sind Wahrscheinlichkeitswerte.

- Entwickelt aus neuesten Forschungsergebnissen. Der zugrundeliegende Algorithmus ist eine Kombination aus Baumkonstruktion, hybridem genetischem Algorithmus und quadratischer Optimierung.
- Die Vorhersagen sind Wahrscheinlichkeitswerte.
- Probleme mit vielen Klassen und Probleme mit zwei Klassen werden in gleicher Weise gelöst.

Versicherungsbeispiel–Fortsetzung

Versicherungsbeispiel–Fortsetzung

Nicht numerische Merkmale müssen umgewandelt werden in numerische.

Versicherungsbeispiel–Fortsetzung

Nicht numerische Merkmale müssen umgewandelt werden in numerische.

Klient	Alter	Geschlecht	verheiratet	Bildung	Kostenklasse
1	25	0	2	2	2
2	45	1	1	0	3
3	32	1	2	3	1
...
100000	21	0	1	1	5

Input-Datei für DATA-FORECASTER

Input-Datei für DATA-FORECASTER

4	Anzahl der Merkmale
100000	Anzahl der Klienten
5	Anzahl der Kostenklassen

Input-Datei für DATA-FORECASTER

4 Anzahl der Merkmale
100000 Anzahl der Klienten
5 Anzahl der Kostenklassen

25	0	2	2	2
45	1	1	0	3
32	1	2	3	1
...
21	0	1	1	5

Vorhersage-Routinen

Vorhersage-Routinen

1. **manuelle.** Merkmale eines Klienten werden in numerischer Form auf dem Bildschirm eingegeben. Vorhersageergebnis wird auf dem Bildschirm wiedergegeben.

Vorhersage-Routinen

1. **manuelle.** Merkmale eines Klienten werden in numerischer Form auf dem Bildschirm eingegeben. Vorhersageergebnis wird auf dem Bildschirm wiedergegeben.
2. **elektronische.** Merkmale vieler Klienten sind in eine Text-Datei geschrieben. Die Vorhersagen für alle Klienten werden in eine zweite Text-Datei geschrieben.

Ergebnis eines Kreuzvergleichs

Ergebnis eines Kreuzvergleichs

Unsicherheit des Modells = 0.63249

Unsicherheit ohne Modell = 1.49802

Eine Trefferquotenmatrix H in % :

Ergebnis eines Kreuzvergleichs

Unsicherheit des Modells = 0.63249

Unsicherheit ohne Modell = 1.49802

Eine Trefferquotenmatrix H in % :

Zeile 1:	61.27	17.68	12.22	5.44	3.40
Zeile 2:	10.98	59.05	18.52	7.67	3.78
Zeile 3:	5.06	14.80	68.33	8.91	2.90
Zeile 4:	1.41	3.93	12.88	76.66	5.12
Zeile 5:	3.23	4.85	10.73	13.29	67.90

Zeile i von H ist die durchschnittlich vorhergesagte Wahrscheinlichkeitsverteilung des Klassenindex für alle Klienten einer Testmenge, die alle Klasse i angehören.

zweites Beispiel: Credit-scoring

Profil eines Klienten:

zweites Beispiel: Credit-scoring

Profil eines Klienten:

- Die Höhe des beantragten Kredits.

zweites Beispiel: Credit-scoring

Profil eines Klienten:

- Die Höhe des beantragten Kredits.
- Anzahl der Beschäftigten der Firma des Klienten.

zweites Beispiel: Credit-scoring

Profil eines Klienten:

- Die Höhe des beantragten Kredits.
- Anzahl der Beschäftigten der Firma des Klienten.
- Jahresumsatz.

zweites Beispiel: Credit-scoring

Profil eines Klienten:

- Die Höhe des beantragten Kredits.
- Anzahl der Beschäftigten der Firma des Klienten.
- Jahresumsatz.
- Höhe der Verbindlichkeiten.

zweites Beispiel: Credit-scoring

Profil eines Klienten:

- Die Höhe des beantragten Kredits.
- Anzahl der Beschäftigten der Firma des Klienten.
- Jahresumsatz.
- Höhe der Verbindlichkeiten.
- Wachstumsrate während der letzten paar Jahre.

- Alter der Firma.

- Alter der Firma.
- Branche, unterteilt in 4 Kategorien:

- Alter der Firma.
- Branche, unterteilt in 4 Kategorien:
 - a) Hochtechnologie,
 - b) Verarbeitendes Gewerbe,
 - c) Transport,
 - d) andere Branchen

Klassifizierung der Klienten in drei Klassen:

Klassifizierung der Klienten in drei Klassen:

1. Der Klient zahlt den Kredit ohne Probleme zurück.

Klassifizierung der Klienten in drei Klassen:

1. Der Klient zahlt den Kredit ohne Probleme zurück.
2. Der Klient zahlt den Kredit mit Problemen zurück.

Klassifizierung der Klienten in drei Klassen:

1. Der Klient zahlt den Kredit ohne Probleme zurück.
2. Der Klient zahlt den Kredit mit Problemen zurück.
3. Der Klient zahlt den Kredit nicht zurück.

Zusammenfassung von DATA-FORECASTER

Zusammenfassung von DATA-FORECASTER

- mathematisch ausgeklügelt, robust, schnell

Zusammenfassung von DATA-FORECASTER

- mathematisch ausgeklügelt, robust, schnell
- Jeder kann es benutzen, einfache Benutzerschnittstelle

Zusammenfassung von DATA-FORECASTER

- mathematisch ausgeklügelt, robust, schnell
- Jeder kann es benutzen, einfache Benutzerschnittstelle
- Anwendung bei Problemen, bei denen Daten analysiert und für Vorhersagen verwendet werden sollen.

Zusammenfassung von DATA-FORECASTER

- mathematisch ausgeklügelt, robust, schnell
- Jeder kann es benutzen, einfache Benutzerschnittstelle
- Anwendung bei Problemen, bei denen Daten analysiert und für Vorhersagen verwendet werden sollen.
- Untersuchung, ob Korrelationen in Daten vorhanden sind oder nicht.

Zusammenfassung von DATA-FORECASTER

- mathematisch ausgeklügelt, robust, schnell
- Jeder kann es benutzen, einfache Benutzerschnittstelle
- Anwendung bei Problemen, bei denen Daten analysiert und für Vorhersagen verwendet werden sollen.
- Untersuchung, ob Korrelationen in Daten vorhanden sind oder nicht.
- Bewertung von anderen Vorhersagemethoden, um Vergleiche von unterschiedlichen Methoden zu ermöglichen.