DATA-FORECASTER

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

April 15, 2003

What is DATA-FORECASTER?

What is DATA-FORECASTER?

- a data-mining software

What is DATA-FORECASTER?

- a data-mining software
- a new unpublished algorithm

What is DATA-FORECASTER?

- a data-mining software
- a new unpublished algorithm
- easy to use, clear user-interface

What is DATA-FORECASTER?

- a data-mining software
- a new unpublished algorithm
- easy to use, clear user-interface
- many applications in
  - ⋆ economics,
  - ⋆ medicine,
  - ⋆ investment banking,

- customer-relationship management,
- insurance,
- fraud-detection,
- pharmacy.

Insurance example

We have data from the past

Insurance example

We have data from the past

client	age	sex	marital status	education	cost class
1	25	male	married	university	2
2	45	female	single	school-drop-out	3
3	32	female	married	PhD	1
...	...	...	...	...	...
100000	21	male	single	high-school	5

Insurance example

We have data from the past

client	age	sex	marital status	education	cost class
1	25	male	married	university	2
2	45	female	single	school-drop-out	3
3	32	female	married	PhD	1
...	...	...	...	...	...
100000	21	male	single	high-school	5

Client arrives, we know his age, sex, marital status, education but not (yet) his cost class.

DATA-FORECASTER returns a forecast for the cost class.

DATA-FORECASTER returns a forecast for the cost class.

	class	1:	1.5 %
	class	2:	77.5 %
For example:	class	3:	9,3 %
	class	4:	6.2 %
	class	5:	5.5 %

Principal tools of DATA-FORECASTER

Principal tools of DATA-FORECASTER

- Construct a forecast model.

Principal tools of DATA-FORECASTER

- Construct a forecast model.
- Forecast to which class an individual belongs.

Principal tools of DATA-FORECASTER

- Construct a forecast model.
- Forecast to which class an individual belongs.
- Measure the quality of a forecast model with cross-validation. i.e. Compare the forecast for the class with the true class of individuals from a test-set from the past.

Principal tools of DATA-FORECASTER

- Construct a forecast model.
- Forecast to which class an individual belongs.
- Measure the quality of a forecast model with cross-validation. i.e. Compare the forecast for the class with the true class of individuals from a test-set from the past.
- Investigate whether the user's data contains correlations.

Principal tools of DATA-FORECASTER

- Construct a forecast model.
- Forecast to which class an individual belongs.
- Measure the quality of a forecast model with cross-validation. i.e. Compare the forecast for the class with the true class of individuals from a test-set from the past.
- Investigate whether the user's data contains correlations.
- Evaluate the quality of any other prediction method, so that it can be compared against DATA-FORECASTER.

Why use DATA-FORECASTER?

Why use DATA-FORECASTER?

- You can do data-mining inhouse, no need to give sensitive data outside your company.

Why use DATA-FORECASTER?

- You can do data-mining inhouse, no need to give sensitive data outside your company.
- It is relatively cheap. No more need to do longer-term projects with data-mining consultants.

Why use DATA-FORECASTER?

- You can do data-mining inhouse, no need to give sensitive data outside your company.
- It is relatively cheap. No more need to do longer-term projects with data-mining consultants.
- It is fast. Computation time for model construction
  - ⋆ 20 seconds for a 1000 X 15 problem
  - ⋆ 37 minutes for a 1200 000 X 15 problem

- Cutting edge research. The underlying method is a combination of tree-construction, hybrid genetic algorithm and quadratic programming.

- Cutting edge research. The underlying method is a combination of tree-construction, hybrid genetic algorithm and quadratic programming.

- The forecasted values are probabilities.

- Cutting edge research. The underlying method is a combination of tree-construction, hybrid genetic algorithm and quadratic programming.
- The forecasted values are probabilities.
- Problems with many classess are treated the same way as problems with two classes.

Insurance example continued

Insurance example continued

Non-numerical characteristics must be converted to numerical ones.

Insurance example continued

Non-numerical characteristics must be converted to numerical ones.

client	age	sex	marital status	education	cost class
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 file for DATA-FORECASTER

Input file for DATA-FORECASTER

4	number of characteristics
100000	number of clients = data records
5	number of cost classes

Input file for DATA-FORECASTER

4	number of characteristics
100000	number of clients = data records
5	number of cost classes

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

Forecast routines

Forecast routines

1. manual. Enter characteristics on the screen. Result is returned on the screen.

Forecast routines

1. manual. Enter characteristics on the screen. Result is returned on the screen.

2. electronic. Write characteristics of many objects in a text-file. All forecasts will be written in a second text-file.

Cross-validation output

Cross-validation output

model-uncertainty = 0.63249

no-model-uncertainty = 1.49802

The hit rate matrix H in % :

Cross-validation output

model-uncertainty = 0.63249

no-model-uncertainty = 1.49802

The hit rate matrix H in % :

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

Row i of H is the average predicted probability distribution for the class index over all test-set data-points who belonged to class i.

Credit-scoring example

Profile of clients:

Credit-scoring example

Profile of clients:

- The credit sum which the client is asking for.

Credit-scoring example

Profile of clients:

- The credit sum which the client is asking for.
- The number of employees of the client’s business.

Credit-scoring example

Profile of clients:

- The credit sum which the client is asking for.
- The number of employees of the client’s business.
- The annual turn-over of the client’s business.

Credit-scoring example

Profile of clients:

- The credit sum which the client is asking for.
- The number of employees of the client’s business.
- The annual turn-over of the client’s business.
- The amount of the current liabilities.

Credit-scoring example

Profile of clients:

- The credit sum which the client is asking for.
- The number of employees of the client’s business.
- The annual turn-over of the client’s business.
- The amount of the current liabilities.
- The growth-rate of the business over the last three years.

- How many years ago the business was founded.

- How many years ago the business was founded.
- The type of business which is divided into four categories:

- How many years ago the business was founded.
- The type of business which is divided into four categories:

a) high-technology,
b) manufacturing,
c) transportation,
d) other services

Classification of clients in three classes:

Classification of clients in three classes:

1. The client payed back the credit without any problems.

Classification of clients in three classes:

1. The client payed back the credit without any problems.
2. The client payed back but with difficulties, special arrangements had to be done.

Classification of clients in three classes:

1. The client payed back the credit without any problems.
2. The client payed back but with difficulties, special arrangements had to be done.
3. The client failed to pay back, the financial institution made a loss.

Summary of DATA-FORECASTER

Summary of DATA-FORECASTER

- mathematically advanced, robust, fast

Summary of DATA-FORECASTER

- mathematically advanced, robust, fast
- anybody can use it, easy and clear user-interface

Summary of DATA-FORECASTER

- mathematically advanced, robust, fast
- anybody can use it, easy and clear user-interface
- can be applied to problems where data has to be analyzed and used for making predictions.

Summary of DATA-FORECASTER

- mathematically advanced, robust, fast
- anybody can use it, easy and clear user-interface
- can be applied to problems where data has to be analyzed and used for making predictions.
- can investigate if there are correlations present in the data.

Summary of DATA-FORECASTER

- mathematically advanced, robust, fast
- anybody can use it, easy and clear user-interface
- can be applied to problems where data has to be analyzed and used for making predictions.
- can investigate if there are correlations present in the data.
- includes a routine for evaluating other forecasting-methods.