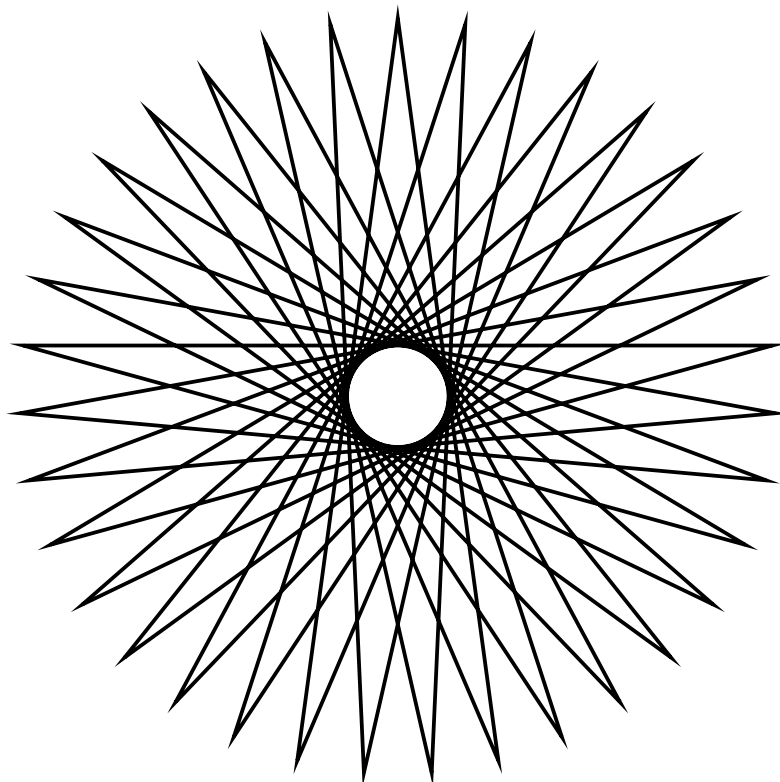


THE DALLAS PROJECT

Report from the NUTEK-supported project AIS-8:
Application of Data Analysis with Learning Systems,
1999–2001



Edited by Anders Holst

Contents

1	Introduction	1
1.1	The DALLAS project	1
1.2	Goals and objectives	1
1.3	Approach and experiences	2
1.4	Participants	2
1.5	This report	3
2	Method descriptions	5
2.1	Regression	5
2.2	Classification	8
2.3	Multilayer Perceptrons and Error Backpropagation	11
2.4	A guide to recurrent neural networks and backpropagation	17
2.5	Inductive Logic Programming	27
2.6	The Bayesian modeling tools	30
2.7	Self-organizing feature maps	36
2.8	Genetic Algorithms	39
2.9	Information theory	41
2.10	Ensembles, Boosting and Bagging	45
3	AstraZeneca: Classification of cancer using 2D-electrophoresis	47
3.1	Introduction	47
3.2	The model used at SICS	48
3.3	The models used at Halmstad University	51
3.4	The model used at Skövde University	54
3.5	Discussion	55
3.6	References	55

4 EKA Chemicals:	
The hydrogen peroxide production process	57
4.1 Introduction to the task	57
4.2 The model used at Halmstad University	63
4.3 The model used at Skövde University	72
4.4 The model used at SICS	80
4.5 The model used at Mitthögskolan	87
4.6 The model used at DSV	94
4.7 Results from the blind test	95
4.8 Summary and discussion	95
4.9 References	96
5 Nordisk Media Analys (NMA):	
The brand awareness task.	97
5.1 Introduction	97
5.2 The models used at SICS	99
5.3 The model used at Mitthögskolan	100
5.4 The model used at DSV	101
5.5 The model used at Skövde University	102
5.6 The model used at Halmstad University	109
5.7 Results	124
6 NovaCast:	
Prediction of alloy parameters	125
6.1 Introduction	125
6.2 The model used at Halmstad University	128
6.3 The model used at SICS	132
7 SCA:	
The dewatering task	135
7.1 Introduction	135
7.2 The model used at SICS	137
7.3 The model used at Halmstad University	140
7.4 The model used at Skövde University	149
7.5 The model used at DSV	154
7.6 The model used at Mitthögskolan	155
7.7 Discussion	158

7.8	References	159
8	Telia:	
	Detection of frauds in a Media-on-Demand system in an IP network	161
8.1	Problem description	161
8.2	Data description	162
8.3	The model used at SICS	164
8.4	The model used at Halmstad University	168
8.5	The model used at DSV	170
8.6	The model used at Mitthögskolan	172
8.7	Discussion	174
8.8	References	175
9	Ericsson:	
	Quality of Service in IP-networks	177
9.1	The task	177
10	Discussion and conclusions	179
10.1	Evaluation of the project	179
10.2	Comments on the project form	180
10.3	Conclusions	180
10.4	Publications made during the project	181

Chapter 1

Introduction

Björn Levin

1.1 The DALLAS project

The DALLAS (“application of Data AnaLysis with LeArning Systems”) project has been designed to bring together groups using learning systems (*e. g.* artificial neural networks, non-linear multi-variate statistics, inductive logic etc) at five universities and research institutes, with seven companies with data analysis tasks from various industrial sectors in Sweden. An objective of the project has been to spread knowledge and the use of learning systems methods for data analysis in industry. Further objectives have been to test the methods on real world problems in order to find strengths and weaknesses in the methods and to inspire research in the area.

1.2 Goals and objectives

Data analysis (*i. e.* the search for and the analysis of structures and dependencies in data) is becoming a more and more important concept in almost any industrial sector. With an ever increasing amount of automated measuring devices, sensors, computerized control equipment, networked accounting systems, internet trade etc, huge amounts of data are collected in any kind of industry or business; data that contain very valuable clues on how to improve the businesses in question. Due to the sheer size, manual analysis of these data sets is virtually impossible. However, despite obvious differences in what is measured in *e. g.* telephone networks, chemical plants, and advertising, the same methods for automated or semi-automated analysis can be applied, and there is therefore a need for similar data analysis tools in a large number of very different industries. A primary goal of the project has therefore been to forward the knowledge about existing new data analysis methods to the industrial partners, to test and show the usefulness of these methods and to establish them as alternatives to existing methods.

Another primary goal has been to supply the academic partners with real world problems and data, industrial feed-back, and inspiration for future research. Such important information, that cannot easily be found in laboratory environments, is of course essential for improving the methods.

The gains of using data analysis tools lie on several levels. Considerable advantages can be gained by simply re-utilizing data collected for various low-level control or administration purposes in a more global analysis. These gains are expected in the form of more even production, lower resource consumption and better competitiveness. Another important gain is better insight into the dependencies and relations in the processes in question, insights that in turn can enable improved production.

1.3 Approach and experiences

A list of tasks (*e. g.* data sets or processes to analyse) was set up by each participating industrial partner defining one or two items. Each industry was then assigned a main academic contact point, and visits arranged for the whole group of academic partners to each industry in order to gather background knowledge. The problems were at that point defined in more detail and *e. g.* formats of transferred data agreed. During the planning of the project it was anticipated that this step would require considerable time and resources, but the time actually needed still exceeded what was expected.

After each industry had collected its data, it was sent to their respective main academic contact points for initial testing and further editing. This turned out very well, since usually several iterations between the industry and the academic contact point were needed and the approach kept the required coordination down to two people.

Once edited, almost all tasks were sent to almost all academic partners and attacked with their favorite methods. A disadvantage was of course that the small resources of the project were divided into even smaller bits by this scheme. The advantage was, on the other hand, that a wider range of methods were tested on the tasks.

The preliminary results were then presented at the industries and refinements in the task definitions or in the collection of the data were decided for a second round of attack.

Finally, in some cases, a competition was arranged between the methods of the academic partners. This was very much appreciated by both the academic partners and the industries in question.

Although many of the learning system methods showed some weaknesses that need to be worked out and although some of the tasks turned out to be too difficult to make real progress on during the project, good and valuable results were obtained for a majority of the tasks and a large amount of insight was gained both among the industrial partners and the academic partners. We feel that the main objectives of the project were fulfilled.

1.4 Participants

The following persons from the five academic and seven industrial partners were involved in this project.

Academic partners

SICS, Swedish Institute of Computer Science, Adaptive Robust Computing laboratory:

Björn Levin (project manager)
Anders Holst
Daniel Gillblad

University of Halmstad, school of Information Science, Computer and Electrical Engineering:

Thorsteinn Rögnvaldsson
Mikael Bodén
Jim Samuelsson

University of Skövde, dept. of Computer Science:

Lars Niklasson
Henrik Jacobsson
Fredrik Linåker
Ulf Johansson

Stockholm University, dept. of Computer and Systems Sciences (DSV):

Lars Asker
Henrik Boström

Mitthögskolan, dept. of Physics and Mathematics:

Mikael Hall
David Martland
Johan Torbiörnson

Industrial partners**AstraZeneca:**

Sven Jacobsson
Anders Hagman
Bo Franzén
Fredrik Andersson

EKA Chemicals:

Lars Renberg
Rolf Edvinsson Albers
Håkan Persson

Ericsson Switchlab:

Harald Brandt

Nordisk Media Analys:

Kristina Ericson
Johan Karlsson
Maria Celén
Helena Aava

NovaCast AB:

Rudolf Sillén
Thomas Karlsson

SCA:

Hans Pettersson
Anders Johansson
Joar Lidén
Göran Sundh

Telia:

Anders Rockström
Rolf Hulthén

1.5 This report

This report has two main parts. The first part is contained in Chapter 2, in which the different methods used in the project are described, both the actual learning system methods and various auxiliary techniques that have been useful. The second part is in Chapter 3 to Chapter 9, containing descriptions of the different industrial applications, and the results achieved when applying different methods to them. Finally, Chapter 10 contains a summary and general conclusions.

