Networks 23 – June 1999

Free Meeting in Bristol
13-14 July 1999

The University of Bristol will host the second NCAF meeting of 1999. The fact that the Joint Meeting in April was run by the IEE means that NCAF incurred very little expense, which leaves us in the enviable position of being able to waive the meeting fee for Bristol!

This meeting also sees the adoption of a new format for the programme. The first day will be devoted to a particular theme, while the second will be composed of the familiar general applications talks. The theme for Bristol is 'Data Mining' or 'Knowledge Discovery in Databases' (KDD). This is a discipline concerned with the efficient recovery of information from large databases and often draws on methods of computational intelligence in order to define effective search algorithms and also to visualise the data.

The first day explores the Data Mining theme and begins (after registration, coffee and the usual welcome), with a tutorial delivered by Simon Cumming of British Airways. Because the subject may not be familiar to all present, this will begin at a fairly elementary level. After morning coffee the session continues with the first of the application talks, the intriguingly titled 'A study of Bogus Official Burglaries Within West Midlands Police' by Inspector Rick Adderley of the said police force. The second talk is by Professor Jim Baldwin of the University of Bristol and is concerned with applications of Artificial Intelligence, the title is 'Machine Learning Methods Using Mass Assignment Theory'.

After lunch, the programme resumes with the invited talk. This is by Kathryn Burn-Thornton of Plymouth University and is concerned both with broad issues of data mining and with specific applications within the experience of her research group. The second talk of this session is by Yike Guo of Imperial College on further applications of Data Mining. Tom Kabaza of SPSS (formerly ISL), will open the final session of the day with a presentation: 'Data Mining with Clementine'. This is concerned with applications of this well-established product. Finally, the day closes with a talk by Professor Vic Rayward Smith of the University of East Anglia, with the self-explanatory title 'Experiences of Data Mining for the Insurance Industry'.

This rather intensive first day will be balanced by a less-demanding second day which will end a little earlier than usual to allow more travel time. The social event this time is a traditional skittles evening at the popular British Aerospace Welfare Association (BAWA), where we have sole use of all three skittle lanes from 7-11pm. There will be a late buffet supper, and (of course) a bar. There will be transportation to and from the university halls.

The second day begins with an extended presentation by Ian Nabney of Aston University on the Neurosat project. This is concerned with the retrieval of sea-surface winds from satellite data. It involves neural networks, Generative Topographic Mapping, Gaussian processes and Bayesian inference to solve a very interesting application and (provably) beat the Meteorological Office

system. After coffee the morning session resumes with a talk by Simon Hickinbotham of the University of York on 'Determining Strain Gauge Faults in Stress Cycle Count Matrices', which is concerned with identifying bad sensor data in aircraft operational load monitoring. The final talk of the morning is by Tshlidizi Marwala of the University of Cambridge on 'Damage Identification in Structures Using Neural Networks and Frequency Response Functions'. This is concerned with the interesting problem of combining the results from several neural nets. Before lunch, Graham Hesketh will be taking us through the solution of the latest Puzzle Corner.

After lunch, the afternoon commences with a presentation from Kevin Bossley of Southampton's Parallel Applications Centre. Entitled 'Neural Network Approaches in Power Station Turbine Monitoring', it discusses a variety of techniques from wavelet packet analysis to Bayesian inference. The meeting closes with an extended presentation from Colin Campbell of the University of Bristol. 'Support Vector Machines: Recent Advances' is concerned with covering the groundwork for these fast-developing systems and also pointing the way for future research by discussing the recently proposed 'Bayes-Point Machines'.

Keith Worden
University of Sheffield

Joint NCAF/IEE Birmingham Meeting Review
Thursday 22 and Friday 23 April 1999

Condition Monitoring: Machinery, external structures and health
Austin Court, The Midlands Engineering Centre, Birmingham

This is the first time that NCAF has visited Austin Court. The venue was in the centre of Birmingham yet with the extensive restored canal system and footpaths between Austin Court and the main hotels there was an almost rural air to the meeting.

The meeting covered a wide range of health monitoring topics with a good balance of applications and methods review presentations. Professor Niranjan set the mood of the meeting with a light hearted, yet rigorous discussion of the merits of Bayesian methodology. An important message was not to hide prejudice or poor frequentist statistics under the label of prior probability estimates. On the other hand, in sequential learning problems where it is undesirable to store the whole data set, Bayes methodology does give a way to continuously update a sound model. We also learned some of the benefits of Bagging (averaging multiple models trained on different randomly sampled data subsets) and Boosting (using a sequence of classifiers each weighting the data which earlier models had difficulty classifying). I had previously thought that Boosting was a dubious if not discredited procedure, but I stand corrected. It is equivalent to working on the convex hull of a set of classifiers. As such, the best classifier performance for a particular false positive rate may be constructed from an appropriately weighted combination of individual classifiers. These points were amply illuminated by practical applications in the field of medical diagnostics.

Real world application

In complete contrast to Professor Niranjan's talk, Chuck Farrar gave a review of

real world application of vibration based structural health monitoring. The methods described are often seen in University Research Laboratories, but in Chuck's presentation they took on a whole new meaning when applied to road bridges. His experiments included measuring vibration response as a function of progressively large defects in bridges across the Rio Grande and the comparison of the results with finite element simulations. As Chuck pointed out, destructive tests on full size structures are comparatively rare.

Professor Tarassenko described another expensive set of experimental tests. He gave a preview of his contribution to the University of Oxford Lubbock Memorial Lectures. He illustrated the need to balance the complexity of the statistical models used to describe normality to the volume of data available. He described how he has been able to build a model of normality for the temperature distribution in the exhaust of a jet engine and how he demonstrated to his customer the need for engine specific models in order to achieve adequate sensitivity to small changes in engine condition. The expensive engineering came in when a sequence of engine tests was performed with deliberately introduced combustion faults to validate the approach used. Finally Professor Tarassenko presented the results of a double blind trial in which his team successfully diagnosed a simultaneous trio of faults which occurred naturally in an engine used to pump natural gas. It is rare to see double blind methodology used outside the life sciences. It should be encouraged as an improvement in the objectivity of an investigation. When the results are finally compared with the hidden diagnosis and found to be accurate, I can confirm from first hand experience that the customer is likely to be very impressed.

Rigorous methods

Professor Tarassenko's paper was complemented by the work of Paul Wells and Dr Steve Roberts (reported by Dr Wright from British Aerospace). This used similarly rigorous methods to validate the location of acoustic emission events in aircraft structures. The contrast was provided by the ability to model the signal propagation behaviour in this case. The presentation described the use of some elegant mathematical simulations to identify the observability of structural defects with different acoustic emission sensor locations. There was a particularly imaginative application of embedding dimension to simplify the characterisation of the very high dimensional signal. Embedding dimension is more usually associated with the analysis of Chaotic systems, but the approach which was seen to be effective, draws another parallel with the non-linear dimensional reduction techniques used by Professor Tarassenko to detect novelty.

The variety of applications areas continued through the rest of the meeting. Professor Patton covered the integration of knowledge based qualitative models with data based quantitative models in control system diagnostics. Dr Abbod described the use of similar knowledge based techniques to the control and diagnosis of faults in drug infusion for administering anaesthetics. Overall, the synergy between the widely differing techniques and applications areas was recognised by most delegates and we left the colloquium happier and wiser than we arrived.

As usual with NCAF meetings, there was an entertaining pantomime enactment of the solution to Networks' last puzzle corner, most ably stage-managed by Graham Hesketh. Graham also organised what turned out to be a rather difficult but highly entertaining quiz for those delegates staying overnight in

Birmingham.

Peter Cowley
Rolls-Royce

Conference Report: PADD'99

Third International Conference on the Practical Applications of Knowledge Discovery and Data Mining

The PADD'99 conference was held at the Commonwealth Institute in London on 21-23 April 1999, as part of the Practical Applications Expo umbrella which also included conferences on knowledge management, Java and constraint logic programming. The conference unfortunately overlapped with the NCAF/IEE two-day meeting in Birmingham.

PADD'99 attracted an international turnout of speakers from as far away as Canada, Brazil, Taiwan, Australia and Martlesham Heath.

Themes which emerged at the conference were around the production use of data mining, ensuring best practices and integrating data mining with On-line Analytical Processing and other elements of data analysis. There were some innovative approaches to data visualisation and some thought-provoking applications.

Setting of expectations

Lisa Sokol, from MRJ Consultancy, in Virginia, USA, spoke on ensuring the success of commercial data mining projects and emphasised the importance of finding the right customer in the organisation, someone looking for new insight into business problems, addressing solvable but critical problems with well targeted objectives. Setting of expectations was important, with each insight evaluated, tested and proven, keying in with customers' knowledge of how to profit from the information. Lisa stressed that we should produce understandable results, exploiting customer knowledge to reduce the search space without ruling out the good portions. Data mining systems should integrate with the customer's infrastructure and expect change, often building in novelty detection models to keep the application up to date.

These sentiments were underlined by high quality contributions from Chris Harris-Jones of consultants AMS, and Barry DeVille of the Canadian office of the supplier Angoss. Barry took a knowledge management stance and discussed best practices, asking how we can go 'beyond data mining' to using customer data to shape strategic direction. He referred to the APQC (American Productivity and Quality Council: http://www.apqc.org, a well-known organisation in knowledge management circles which provides resources on 'best practices' of all kinds) and has begun to be active in best practices for data mining and is carrying out some benchmarking studies, and to CRISP-DM (see http://www.ncr.dk/crisp), a European project to develop a cross-industry standard process for data mining. In his talk, Chris addressed how data mining and text mining fit with emerging knowledge management technologies and whether a data warehouse is really necessary. After mining has built a model, what next? Model implementation and 'incremental mining' were recurring themes. Chris touched on the automation of data mining of web transactions

and on implementation and incremental improvement by the 'champion/ challenger' framework, i.e. always having at least a second model in the frame. Discussion involved ideas of partially automating data mining by means of application templates or wizards, where experience of how best to go about particular vertical applications is embodied.

First-cut filtering

Visualisation was emphasised, and Paul Coker of the Future Technologies Group in BT Labs (http://www.labs.bt.com) showed a 'data chromatography' scheme based on the idea of chromatography in the physical sciences, used for separating mixtures. The user poses questions and positions them on the screen to act as 'magnets' pulling the customers' representation dots towards them. This can act as a first-cut filtering method for potentially billions of telephone traffic records. Ming Hao, from Hewlett Packard Research Labs in the USA, visualised data using a hyperbolic space, which could be configured to show a tree or multiple connections between variables.

Finally, a couple of applications; Warren Graco, from the Health Insurance Commission in Australia, used data mining to identify 'doctor-shoppers', drug-dependent patients who shop around; Ira Moskowitz, from the US Naval Research Laboratory, gave a fascinating talk on 'The Rational Downgrader', where the issue is one of 'data damming', i.e. making sure that people cannot use data mining to recreate sensitive data from what is made publicly available. This is done by inserting missing values in the right place. (Sounds very similar to some of our 'Rottweiler' puzzles!)

Of the talks I have not described, there will be a second chance to hear two of them, one by Kathryn Burn-Thornton of Plymouth, and one on data mining at British Airways, at the NCAF meeting in Bristol in July. See you there!

Simon Cumming
British Airways plc

PUZZLE CORNER Number 9

The gothic towers of the asylum rose stark against the crimson sunset sky. Lisa entered the Wendleberry Home for the Academically Challenged Oxbridge Scholars with some trepidation. Maybe she had been a bit too generous in her support for the social rehabilitation programme.

She approached the bed of one wizened individual and noticed a beautiful multicoloured walking stick at his side. It was constructed of 11 sections, each a different colour.

'I bought that when I was 12 years old from a Chinese curiosity shop', croaked the old man. 'Each piece is an exact number of inches long, and they are all interchangeable except for the handle piece. The owner said to me, "Don't get it wet, don't feed it after midnight, and there is one configuration of the pieces which will let you use the stick as an accurate measure for any length, in whole inches, from 1 to 48." I have tried different combinations every morning, noon and night since then and I have never found that elusive arrangement. I'm 100 tomorrow, and I feel that I have only a few more attempts left in me.'

As the old man drifted off to sleep, Lisa set about re-assembling the walking stick for the last time. When the centennial man woke in the morning he could not believe his eyes. The perfect walking stick of his dreams was there beside his bed. He couldn't contain himself. 'Damn', he said, 'I wanted to do that myself!' Lisa replied, 'Get a life.'

Given that the handle piece was 3 inches long, and the other pieces were 8,8,8,8,4,4,3,3,1 and 1 inches long, what was the arrangement which Lisa discovered? Also, what was the oldest the man could possibly have been before he must have lost his combinatorial faculties? The answers will be given at the next NCAF meeting (13-14 July 1999, Bristol University).

Fenella the Rottweiler

Enterprise Miner - a personal view

SAS has established itself as one of the leading suppliers of what it terms 'information delivery' software products and is expanding from its well-established base of statistical software into data warehousing, web enablement, OLAP and 'vertical market' applications.

Internationally, SAS has been following developments in neural networks and data mining technologies for several years now, and Warren Sarle, at their headquarters in Cary, North Carolina, in the USA, will be known to many people as the maintainer of the neural nets 'FAQ' (frequently-asked questions) list on the Internet. His style of healthy scepticism and strong adherence to statistical principles has focused the direction in the application of neural networks. Provision by SAS for data mining using neural networks was originally in the form of a set of macros written by Warren Sarle and using standard non-linear programming functionality in the SAS/OR module, treating neural nets as optimised parameter models. Also for a number of years, tools such as SI-Chaid from third party suppliers have been available to run on SAS for development of decision trees.

Add-on product

Enterprise Miner is a data mining add-on product, which requires some other SAS modules to be in place. It is an exciting development from SAS, in that it supports their methodology for data mining, which SAS calls SEMMA (standing for Sample, Explore, Modify, Model, Assess). The user interface to Enterprise Miner has a visual-programming style (in a similar way to Clementine, for example), which makes the methodology clear, illustrating the flow of control and data, and takes care of organising work into 'projects' and 'diagrams'. This way, the many different versions of models that arise in a data mining exercise are automatically organised in a structured way. Good practices such as preliminary understanding of the data, pre-processing, partitioning the data into training, test and validation sets, and treatment of missing values, are well supported.

Enterprise Miner supports decision tree, neural net and association rule methods, as well as access to traditional, statistical methods such as clustering and regression. With the decision trees, chi-squared, entropy and Gini reduction methods are supported, and in the case of neural nets, there is a choice of a number of training schemes including Levenberg-Marquardt, quasi-Newton and conjugate gradient, with a wide range of error functions to suit different

distributions. One feature which seems to be unique is Enterprise Miner's preliminary optimisation, which attempts to find good starting values for the network weights. In terms of model assessment, again the functionality is rich, supporting comparison in terms of measures which take into account model complexity (Aka Iki Information Criterion and Schwarz's Bayesian Criterion), and in terms of profit-related lift charts.

Although strong on functionality and theoretical underpinning, the limitations of Enterprise Miner are some clumsiness in the user interface and an apparently insatiable hunger for machine resources. An example of the former is the print function in the decision tree module, where for anything other than a trivial model one has to resort to a pencil and paper! Enterprise Miner can run in client-server mode, but still has quite heavy requirements on the client machine (minimum memory recommended is 48M), and can be quite profligate in creating internal and temporary files. Something else to watch out for is the terminology, which does not necessarily correspond to that used in other neural network tools.

Having said this, the software is certainly suitable as a powerful data-mining tool for a large company with existing SAS experience. A new version is due out shortly, which will include self-organising maps and improvements to the user interface. For further information contact SAS Institute at Marlow on 01628 486933.

Simon Cumming
British Airways

Diary Dates

6-9 July, COMADEM. Condition Monitoring and Diagnostic Engineering Management, Sunderland. http://www.comadem99.sunderland.ac.uk

10-16 July, IJCNN'99. 10th International Joint Conference on Neural Networks in Washington, DC, USA.
http://www.cas.american.edu/~medsker/ijcnn99/public_html

13-14 July, NCAF Meeting. Bristol University.

Contact: Sally Francis. Tel: +44 1784 477271 or +44 1784 431341 ext 270, fax: +44 1784 472879, e-mail: ncafsec@brunel.ac.uk

9-11 August, Symposium on Intelligent Data Analysis. Amsterdam, NL.
http://www.wi.leidenuniv.nl/~ida99/

9-12 August, AI & Soft Computing. Hawaii, USA.
http://www.iasted.com/conferences/1999/hawaii/asc.htm

23-25 August, Neural Networks for Signal Processing. Madison, Wisconsin, USA. http://eivind.imm.dtu.dk/nnsp99/

1-3 September, Irish Conference on AI and Cognitive Science. Cork.
http://www.cs.ucc.ie/aics99/

3-5 September, European workshop on Neuromorphic

Systems. Stirling. http://www.cs.stir.ac.uk/EWNS2/

7-10 September, ICANN99. International Conference on Artificial Neural Networks, Edinburgh. http://www.iee.org.uk/Conf/ICANN/

28-29 September, NCAF Meeting. Fitzwilliam College, Cambridge.

Contact: Sally Francis. Tel: +44 1784 477271 or +44 1784 431341 ext 270, fax: +44 1784 472879, e-mail: ncafsec@brunel.ac.uk

Members' news and views

Deadline for the next edition is 29 July 1999.

Next Edition

Review of the Bristol Meeting.
Preview of the Autumn Meeting in Cambridge.