ncaf
Natural Computing Applications Forum

Networks

20 years of natural computing
1990–2010

Making Sense of Data – Theory and Practice

University of Surrey, Guildford, 12–13 July 2010

We can extract and store the data – but what do we do with all the giga- or peta- or zettabytes?

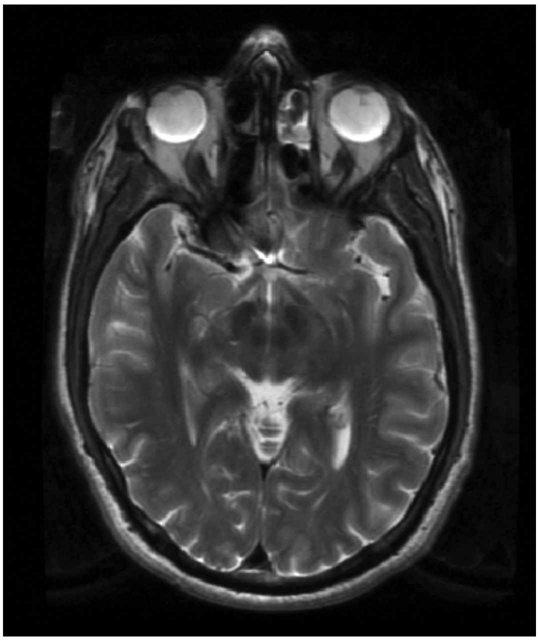
The Department of Computing at the University of Surrey looks forward to hosting the next NCAF meeting on its Stag Hill campus, within walking distance of both Guildford town centre and its railway station.

We are a small, friendly department of around 16 academics, and about half of us are involved in the active research group BIMA (Biologically Modelling and Applications). Although this is the first time we will be hosting the meeting, NCAF members have been involved in past BIMA meetings. One current lecturer even found his PhD topic through an NCAF meeting some years ago!

'Making Sense of Data – Theory and Practice', the topic for the upcoming meeting, reflects the need to develop theory and tools to analyse and categorise the ever-growing amount of raw data that we can extract from natural and artificial systems. We can extract and store the data – but what do we do with all the giga- or peta- or zettabytes? Which parts of it are really significant? Which are useful for the purposes we have in mind? How can we distil the relevant information, often only a few bytes? These are challenging problems that recur in a vast range of current research and applications.

Research in neuroscience has come to a point where ever more data can be collected from hundreds of simultaneous neuron recordings and ever more detailed brain imaging, but the existing tools to analyse this data are at their limits. What is the neural code? Which patterns of neural activity correspond to behaviour? Can we build man-machine interfaces that utilise neural activity for prosthetics or applications in military and gaming applications? Another example is applied research to extract relevant information from video (from CCTV and other recordings used to trace movement) and its significance for the user. Finally, there are also applications in data mining the world-wide web with the aim of creating the 'semantic web'.

We hope that the talks at the NCAF meeting at Surrey will cover all these aspects of understanding data.



The Chairman's brain (from fMRI scan at NCAF meeting,York May 2005).

The themed first day will feature a talk about tracing surgical instruments for the recording of eye operations to improve surgeons' performance. Further papers will outline the use of natural data mining techniques in military applications and how understanding hierarchical processing of visual data in the cortex can help to extract features from video.

On the second day, the topics will include 'analysing jet engines with the help of belief networks' and 'how to use GPGPU (General Purpose Computing on GPUs)' – a low cost approach to parallel processing – in order to simulate spiking neural networks for olfactory processing.

So, why not come along to learn how to make your belief networks fly or how to make your GPUs sniff like an invertebrate at Surrey in July!

As always, watch out for the final programme and the registration details at www.ncaf.org.uk.

Andre Gruning
University of Surrey

The emerging behaviour of complex networks

There are more networks with 17 nodes than atoms on earth.

The theme of the January meeting at Aston University was Complex Systems. The keynote talk was by David Saad (Aston University) who focussed on research to understand the emerging behaviour (macroscopic properties) of complex networks. What makes a complex network? It is large-scale, non-linear, heterogeneous (i.e. both sparsely and densely connected nodes) and hierarchical. The challenges include optimising, managing and controlling such networks. The applications of this work are to telecommunications, the internet, energy networks. David showed how methods from statistical physics and Bayesian inference (particularly belief propagation) can be used to better understand and analyse both specific networks and their general properties.

This theme was continued in the talk by Alexander Stepanenko (Aston University). He discussed in some detail the modelling of large-scale packet-switched Internet-like networks and the theoretical basis for a new generation of routing protocols. He and his collaborators have developed a complete statistical-physics based description of these networks and used this to develop a statistical description of losses in a single buffer and the whole network.

Reimer Kuehn (King's College, London) spoke about risk modeling in financial markets: portfolios, credit, and liquidity. The complexity comes from the functional and dynamic nature of relationships. Contrary to the normal assumptions, risks are not independent or statistically uncorrelated, so conventional models underestimate the probability of large losses by several orders of magnitude. Instead, treating the system as a network of processes on a random graph enables us to build more realistic risk models.

Sparse networks

Sach Mukherjee (University of Warwick) showed how sparse networks were relevant to bioinformatics, network biology and its application to understanding cancer. The link is through protein signalling: aberrant functioning of these networks is implicated in almost every aspect of cancer biology. By understanding the networks better, we can hope to develop rational, targeted therapies to replace the drastic treatments (radiation, chemotherapy) used today. The computational task is to learn these networks from noisy data: they are modelled using Bayesian belief networks, where the directed links are a natural way to denote signalling. This task is particularly challenging since the number of proteins is very large, while the number of observations is relatively small. Hence model and variable selection in a Bayesian framework are critical to making progress. Exact methods are hopeless: the number of possible networks grows super-exponentially. There are more networks with 17 nodes than atoms on earth. So we need strongly sparse priors and sampling methods to explore the posterior distribution over graphs. This approach has been successfully applied to some cancer sub-types and validated biologically.

Jort van Mourik (Aston University) described his recent work in optimisation using particle swarms. By adding periodic dispersion to the basic algorithm it is possible to search effectively on multi-model problems. This improves the solution quality and robustness of the algorithm. Ian Nabney (Aston University) showed how probabilistic methods could be used to measure the complexity of time series. This has the advantage over standard dynamical-systems measures (such as Lyapunov exponents and entropy) of being much more robust to noise. Another benefit is the opportunity to introduce Bayesian methods to determine algorithm parameters and make the process more automated and less dependent on expert intervention. Applications to electrocardiogram signals showed a significant improvement in diagnostic performance compared with current methods.

Computer art

On the second day we had some more general papers. Aniko Ekart (Aston University) demonstrated her program to create abstract computer art. It uses genetic programming to evolve randomly generated images into aesthetic images with symmetries. One function is used to code the entire image by computing the RGB values at each pixel. Both Cartesian and polar coordinates can be used, leading to mirror or radial symmetric components. Because each image is evaluated by the human user (for the fitness function), the population and number of generations are quite small. Nevertheless, it takes 10-15 minutes to obtain the final image. She is now trying to automate the fitness function computation by looking at the frequency of component function usage and constant ranges in particularly nice final images.

Peter Tino (Birmingham University) gave a talk with the intriguing title of 'One-shot learning of Poisson distributions'. He showed that the Audic-Claverie method, which is used to assess the reliability of gene expression profiles on extremely small amounts of data, is sound.

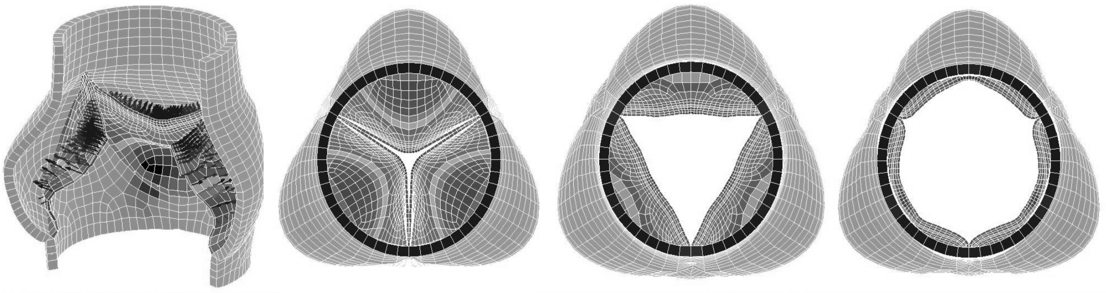
Thorsten Schnier's (Birmingham University) talk was about the use of natural computation to make sense of smart metering data. Because of the need to reduce energy usage, smart metering is being rolled out to domestic properties so that we can understand (and eventually control) our energy consumption. The key challenge is to disaggregate the information: from a single measurement work out what sort of appliances are switched on. The approach that he has taken is to evolve device models (e.g. for domestic appliances of different types, heating etc.) and then fit them to the measurements as a mixture model.

Martin Schroeder (Aston University) showed how visualisation methods can be applied to geochemical data for use in oil exploration. The novelty was the use of block covariances in order to capture correlations between variables. Finally, Maria Chli (Aston University) spoke about the use of agent-based simulations in markets and social systems. She gave three examples: simulations to understand whether segregation of ethnic groups is caused by racism; analysis of the UK Fair Trade market (at £880M it is the largest in the world); and supply chain formation.

The two-day NCAF event comprised the usual mixture of excellent presentations and informal discussion, as well as an important NCAF AGM, where it was decided to hold the next meeting in Guildford.

Ian Nabney
Aston University

Bayesian Sensitivity Analysis



Will Becker

Figure 1: A finite element model of a human heart valve. Left to right: a cross-section showing stresses in the leaflets, and the valve in the closed, half-closed and open positions.

"True wisdom consists in knowing that you know nothing." So said Bill S. Preston Esq. in the seminal work 'Bill and Ted's Excellent Adventure' as he mused over the teachings of Socrates. The root of the philosophy of uncertainty analysis is comparable, although in most cases we hope to know a little more than absolutely nothing. Nevertheless, we wish at least to acknowledge our lack of knowledge, and most importantly, to quantify its effect.

Engineering, science and economics are some of the many disciplines that make use of complex mathematical models – generally solved by computers – to predict the behaviour of a wide variety of problems. But, what happens if there is some doubt about some of the model inputs, the parameters, or even the model itself? A logical prognosis to this problem would be to ask, "is the model output affected?", "by how much?", and finally "what can I do about it?".

Uncertainty analysis addresses the first two issues. *Sensitivity analysis* goes further, and allows us to find the input parameters that are contributing the most (or the least) to the output uncertainty. The various motivations for this include identifying particularly sensitive parameters that can be scrutinised and found to a greater degree of accuracy; conversely, insensitive parameters can be discounted from further study. Thus, the output uncertainty can hopefully be decreased.

Many approaches exist to perform uncertainty analysis, broadly divided by their method of quantification of uncertainty, and level of sophistication. Probabilistic methods are generally well-understood and provide reliable results provided that there is enough information about input uncertainty (see Keith Worden's summary of uncertainty in the Nov 09 newsletter). In this approach, the input $\mathbf{x}$ is treated as a random variable, assigned a probability distribution, and the uncertainty is propagated through the system to provide distributions for model outputs. The output variance (and conditional variance, see later) is an easily interpretable measure of model uncertainty.

A long-established method for propagating probabilistic uncertainty through a system is the use of a Monte Carlo simulation. Samples are drawn from a specified input distribution, and the model is run for each input point and the output distributions can be constructed by building histograms. This is a reliable method, but the obvious drawback is that even for a model of very few input dimensions, the number of sample points (and corresponding model runs) is necessarily very large to gain any kind of accurate output distributions. This is a particular problem when a single run of a model requires a significant amount of time.

Metamodel

One solution to this problem is to build an emulator of the model, otherwise known as a metamodel. The emulator is built from a much smaller number of model runs than the Monte Carlo analysis would require, and once constructed, is computationally vastly cheaper to interrogate than the real model. If (and this can be a significant 'if') the emulator accurately models the response of the real model, then uncertainty and sensitivity data can be calculated for a greatly reduced computational cost.

Gaussian processes were first introduced to modelling computer code by Sacks *et al* in 1989 in a paper entitled 'Design and Analysis of Computer Experiments'. They are particularly suited to dealing with uncertainty since they are parsimonous and return a Gaussian distribution for any input, rather than a crisp value. A prior mean function (often equal to zero or a simple linear fit to the training data) and a prior covariance function, (that represents the belief about the 'smoothness' of the function) are specified. Both functions contain prior distributions of hyperparameters that are estimated by the MAP point or marginalised using MCMC. Our prior functions are then conditioned on the training data, resulting in a posterior mean and covariance function that specify a mean and covariance for any given input.

From our Gaussian process emulator, some very useful uncertainty and sensitivity quantities can be calculated. Of immediate interest are simply the mean and variance of the output. This can be readily estimated from the emulator as the expected value of the posterior mean function, an integral that can be evaluated analytically for common input distributions and covariance functions. The variance can be found in a similar manner. These quantities differ from the sample estimates, since they are integrated over the entire input space, and should represent more accurate estimates, assuming the model provides a good fit.

A well-used quantity in sensitivity analysis is the *main effect*, which is defined as $E(Y|X_i)$, i.e. the expected value of the output Y conditional on a particular input x_i. Again, the posterior expected value of this quantity can be inferred from the posterior mean, this time by integrating over all other input dimensions except *i*. Since this is a function of x_i, an illustrative plot can be created of the effect of varying this input, averaged over uncertainties in other parameters, distinct from the simple effect of varying that parameter alone.

Sets of inputs

In order to quantify the sensitivity of the output to (sets of) inputs, a variance-based approach is adopted. The *main effect* index is defined as the

Continued overleaf

PUZZLE CORNER

Number 43

Down at their local pub, Lisa and 'Slippery Sam' had been engaging in some harmless games of cribbage to while away the long, cold British summer evenings. For Sam, the game was becoming stale, so he suggested a new game based on one of the founding principles of cribbage, namely getting cards to add up to 15.

He suggested a simple game whereby 9 cards, say the Ace to Nine of Hearts, were placed face up on the table. Players would then take it in turns to pick up a card from the table to add to their hand. The winner would be the first player to be able to collect 3 cards which added up to 15 (the Ace counting as 1). If neither had won by the time all 9 cards had been picked up then it was a draw and they could try again, alternating who went first.

Lisa countered by suggesting a more interesting variant whereby the first player to hold 3 cards which added to 15 would be the loser. And, for a final option, she suggested adding the twist that the Ace could be counted either 1 or 10. This now gave four different games to try: (a) Make 15, Ace is 1; (b) Avoid 15, Ace is 1; (c) Make 15, Ace is 1 or 10; (d) Avoid 15, Ace is 1 or 10.

Which, if any, of these games are fair? For those that aren't fair, what is the winning strategy?

The answer will be given at the next NCAF meeting (12/13 July 2010, University of Surrey).

Fenella the Rottweiler

ncaf
Natural Computing Applications Forum

COMMITTEE NOTES

Chairman/Treasurer
Graham Hesketh
Rolls-Royce plc

Secretary
Jo Thomas
Rolls-Royce plc

Editor of Neural Computing & Applications Journal
Professor John MacIntyre
University of Sunderland

Managing Editor of *Networks*
Dr James Hensman
University of Sheffield

Dr David Clifton
University of Oxford

Dr Mark Ebden
University of Oxford

Professor Ian Nabney
University of Aston

Dr Andy Pryke
The Data Mine Ltd

Dr Rajesh Ransing
Swansea University

Please contact NCAF through
Graham Hesketh, Chairman – NCAF
PO Box 5944
Derby DE24 8ZD U.K.
Tel: +44 (0) 1332 240470
Fax: +44 (0) 1332 243530
e-mail: enquiries@ncaf.org.uk
http://www.ncaf.org.uk

MEMBERS' NEWS AND VIEWS

Please send contributions to the Managing Editor – Dr James Hensman, e-mail: James.Hensman@sheffield.ac.uk

Edited and Produced by:
Forum Public Relations
Westgate House, Old Ivy Lane
West End, Southampton
Hampshire SO30 3RX
Tel/Fax: 023 8013 9805
e-mail: info@forum-pr.co.uk

Looking into the future

2010 sees NCAF celebrate 20 years of natural computing!

The meeting at Aston University in January incorporated NCAF's Annual General Meeting and this gave rise to much discussion about the organisation's current state and its future direction.

The Chairman reported that although NCAF is still financially healthy, its income had been reduced due to a fall in attendance levels at recent meetings; however, with a mailing list of over 300 people, NCAF had managed to continue to provide high quality meetings for its members and to maintain a good identity within both academia and industry.

The Chairman asked attendees to consider whether the aims and objectives of NCAF were the same as when the organisation was originally set up and whether its current activities were still relevant.

After some discussion, *complexity* was added to NCAF's scope. Various ideas were also mooted as to the content and format of future meetings. In the past, NCAF has had success with panel discussions and break-out groups. It was felt that having workshops with general discussions targeted at particular topics and real industrial problems could encourage more people to attend and contribute ideas. The NCAF social event is definitely here to stay though – the general feeling was that this is a very important part of every meeting and should be kept!

The need for a publicity officer to join the Committee was also discussed, since there was much interest in promoting upcoming meetings around NCAF's 20th anniversary. If you have some spare time and would like to take up this role then please contact an existing Committee member.

The full minutes of the meeting are available online at www.ncaf.org.uk. However, should you have any further suggestions as to the future of NCAF, then these would be very welcome, particularly at the discussion group uk.groups.yahoo.com/group/ncaforum/, or by direct contact with the Chairman.

Notes from the Chair

It's hard to believe, but we are very close to entering NCAF's third decade of operation. September 1990 marked our first meeting, so the next meeting in January 2011 will celebrate over 20 years of service to the natural computing community. There have been many contributors to the success and longevity of the Forum and it would be really great to reunite as many of these as we can to acknowledge this significant milestone and review the impact of our unique organisation over the last 20 years.

The Chairman

Bayesian Sensitivity Analysis – *continued from page 3*

variance of the main effect, corresponding to the reduction in output variance that would be expected if the true value of x_i were to become known. This quantity can be extended to sets of inputs, representing the effect of the interaction of the uncertainty in a set of parameters, additional to the main effects of each. Thus the total output variance can be systematically broken down into portions representing the contribution from each parameter and all permutations between them. Since it may be tedious to calculate all possible permutations, the *total effect index* is a further convenient measure that sums the variance contributed by a parameter and all interactions associated with it, defined on the complement of *i*.

A brief example shows this methodology applied to a finite-element model of the human aortic heart valve (see Figure 1). The simulation of biomechanical systems is riddled with problems – almost all inputs to the model, such as loading, dimensions and material properties vary significantly from one individual to the next. Additionally, biological tissue is highly nonlinear, anisotropic and heterogeneous.

Uncertainties considered in this model included material properties of the various regions of the valve, such as the leaflets (the moving 'flaps') and the sinus (the rest of the valve). One finding was that when considering stresses in the model, that the stiffness of the sinus was significantly more influential in causing stress uncertainty in the leaflets than the stiffness of the leaflets themselves. From this it could be concluded that the expansion of the sinus is instrumental in the opening of the natural valve, since it allows the leaflets to open without significant buckling. The buckling of the leaflets has long been thought to be responsible for the poor longevity of bioprosthetic valves. Overall uncertainty in the model outputs was high, even for the small parameter set considered.

This example hopefully illustrates that sensitivity analysis can also provide a deeper understanding of a complex model, which may help to make more informed predictions from simulations. Additionally, results from models can be put into context so that they can be used with appropriate caution. After all, as Bill S. Preston Esq. also commented on a conversation with his future self: *why would we lie to ourselves?*

Will Becker
University of Sheffield

DIARY DATES 2010/11

15–18 September – ICANN'10, The 20th International Conference on Artificial Neural Networks, Thessaloniki, Greece.
http://delab.csd.auth.gr/icann2010/

28–30 September – ICNN 2010, International Conference on Neural Networks, Amsterdam, The Netherlands.
http://www.waset.org/conferences/2010/amsterdam/icnn/

6–9 December – NIPS 2010, Neural Information Processing Systems, Vancouver, B.C., Canada. Workshops 10-11 December, Whistler.
http://nips.cc/

January 2011 – NCAF Meeting – Venue and theme TBA. For information email enquiries@ncaf.org.uk or telephone 01332 240470.