Australian Machine Learning Workshop

hosted by the

Machine Learning Group
Research School of Information Sciences and Engineering
Australian National University
Canberra

Seminar Room A105, RSISE Building, ANU
22–23 November, 1999.

Monday, 22 November

9:20–9:30	**Welcome**
9:30–10:30	**Invited speaker:** *M. J. D. Powell, University of Cambridge* On interpolation by radial basis functions
10:30–11:00	**Break**
11:00–12:40	The case for maximum likelihood as a universal training procedure *G. Goodman and G.N. Newsam* *Defence Science and Technology Organisation*
	On the effect of data set size on bias and variance in classification learning *Geoff Webb and Damien Brain, Deakin University*
	Circling the square: networks with hidden features delineated by conic sections *Alan Blair, University of Melbourne*
	On learning decision lists *Douglas Newlands and Geoff Webb, Deakin University*
	A minimum encoding inference approach to theoretical syntax *Mike Dowman and Jon Patrick, University of Sydney*
12:40–2:00	**Lunch**
2:00–3:00	**Invited speaker:** *Claude Sammut, University of New South Wales* Strong typing versus declarative bias in ILP
3:00–3:30	**Break**
3:30–5:10	Learning safe programs *Eric Martin, University of New South Wales*
	Classification of individuals with complex structure *John Lloyd, Australian National University*
	Use of ordinals to model mistake bounds in learnability of logic programs *Arun Sharma, University of New South Wales*
	An inference process for identifying structure in text streams *Jon Patrick and Hong Liang Qiao, University of Sydney*
	The semantic interpretation of discourse (SID) *Stephen Anthony, University of Sydney*

Tuesday, 23 November

9:30–10:30 **Invited speaker:** *Yoav Freund, AT&T Labs – Research*
Decision trees, margins and Brownian motion

10:30–11:00 **Break**

11:00–12:40 Sparse kernel feature analysis
Alex Smola, Australian National University

Maximum margin learning vector quantization
Lawrence Buckingham and Shlomo Geva
Queensland University of Technology

Reinforcement learning in continuous action space: a bicephal solution
Frederic Maire, Queensland University of Technology

Direct gradient-based reinforcement learning
Jonathan Baxter, Australian National University

Reinforcement learning in networks of spiking neurons
Peter Bartlett, Australian National University

12:40–2:00 **Lunch**

2:00–3:00 Computational models of the basal ganglia and cerebellum for sensorimotor control
Marwan Jabri, University of Sydney

Selective attention adaptive resonance theory
Peter Lozo, Defence Science and Technology Organisation

Clustering to enhance FSA extraction from recurrent networks
Dylan Muir, M. Towsey and Joachim Diederich
Queensland University of Technology

3:00–3:30 **Break**

3:30–4:30 Data sharpening
Peter Hall, Australian National University

Smoothing and approximation techniques applied to data mining problems
Stephen Roberts, Australian National University

The effects of class size on template matching
Hari Koesmarno and Warwick Graco,
Health Insurance Commission

4:30 **Close**

Abstracts

On interpolation by radial basis functions

M. J. D. Powell, University of Cambridge

Radial basis functions are highly suitable for approximation in data mining applications, because it is easy to allow large numbers of variables, and because they take advantage automatically of the situation when the data are on a low dimensional manifold in a space of very many variables. Therefore this approach to approximation will be described and discussed briefly. Most of the known theoretical properties of the approach have been discovered by studying interpolation. We will note the accuracy that can be achieved when interpolating to values of a smooth function. Least squares fitting will also be addressed, because it provides some smoothing and can reduce greatly the number of terms that occur in the approximation.

The case for maximum likelihood as a universal training procedure

G. Goodman and G.N. Newsam (Defence Science and Technology Organisation)

Many machine learning problems reduce to training a classifier. Given a specified metric under which the operational performance of the classifier is to be assessed, the natural approach to training is to find the classifier settings that maximise classification performance under this metric on the training data. We present results and arguments that indicate this approach may not in fact be optimal: universal training algorithms may exist that are most likely to identify the best classifier regardless of how performance is subsequently measured.

In particular, in cases where training is equivalent to estimating parameters defining distributions, we show that, under some reasonable assumptions, the parameters determined by training to maximise likelihood will also maximise the subsequent expected performance of the classifier under any reasonable metric. In particular we compare the results of training to maximise likelihood with the results of training to maximise the discriminative likelihood (i.e. training to maximise the expected log probability of correct classification): the latter is a natural performance metric in many problems. We will present numerical experiments comparing the results of training under these two metrics for problems that satisfy the assumptions noted above, and for problems that don't.

On the effect of data set size on bias and variance in classification learning

Geoff Webb and Damien Brain (Deakin University)

Data mining has elevated machine learning to an important role in modern business information technology. However, machine learning evolved in a different research context to that in which it now finds itself employed. A particularly important problem in the data mining world is working effectively with large data sets. However, most machine learning research has been conducted in the context of learning from very small data sets. To date most approaches to scaling up machine learning to large data sets have attempted to modify existing algorithms to deal with large data sets in a more computationally efficient and effective manner. But is this necessarily the best method? This paper explores the possibility of designing algorithms specifically for large data sets. Specifically, the paper looks at how increasing data set size affects bias and variance error decompositions for classification algorithms. Preliminary results of experiments to determine these effects are presented, showing that, as hypothesised, variance can be expected to decrease as training set size increases. No clear effect of training set size on bias was observed. These results have profound implications for data mining from large data sets, indicating that developing effective learning algorithms for large data sets is not simply a matter of finding computationally efficient variants of existing learning algorithms.

Circling the square: networks with hidden features delineated by conic sections

Alan Blair (University of Melbourne)

Two of the most popular connectionist architectures are multi-layer sigmoidal networks and radial basis function (RBF) networks. RBF networks generalize well in situations requiring curved decision boundaries, but do not scale well to high dimensions. Sigmoidal networks have good scaling properties but perform in a piecemeal fashion on problems requiring curved boundaries. This talk will describe a new model which aims to combine the advantages of sigmoidal and RBF networks. In this model, called the neural nutshell, the hidden features may be delineated by any kind of conic section.

On learning decision lists

Douglas Newlands and Geoff Webb (Deakin University)

A decision list is an ordered set of decision rules. Most decision list induction algorithms utilise a variant of the AQ covering algorithm. This approach finds the single most effective rule for the training cases, places it at the head of the decision list, removes the cases that the rule covers, and then repeats the process for each subsequent position in the list until no further rules can be found. Usually, a default rule is added to the end of the list, classifying as belonging to the most common class any cases not classified by an earlier rule. A number of alternative algorithms have emerged, however, that start with the default rule and then add rules into the decision list in front of the default rule [Van Horn & Martinez, 1993; Webb & Brkic, 1993; Cohen, 1995].

These algorithms can provide greater computational efficiency as they deal with smaller training sets than the traditional covering approach. They also typically create sigificantly fewer rules. This paper presents an abstract description of this class of algorithms and outlines their relative strengths and weaknesses.

References

Cohen, William W. (1995) Fast Effective Rule Induction. Proceedings of the Twelfth International Conference on Machine Learning, Morgan Kaufmann.

Van Horn, Kevin S. and Martinez, Tony R. (1993) The BBG Rule Induction Algorithm. AI'93 - Proceedings of the Sixth Australian Joint Conference on Artificial Intelligence. World Scientific. pp. 348-355

Webb, Geoffrey I. and Brkic, Nenad (1993) Learning Decision Lists by Prepending Inferred Rules. Proceedings of the AI'93 Workshop on Machine Learning and Hybrid Systems. Melbourne, pp. 6-10.

A minimum encoding inference approach to theoretical syntax

Mike Dowman and Jon Patrick (University of Sydney)

Chomsky (1986) has influentially argued that syntactic theories should be psychological theories of a person's knowledge of language, and as such they should be able to account for how children acquire language. Chomsky has proposed that the structure of all languages is determined by an innate universal grammar, and that variation between the syntactic systems of different languages can be accounted for by the setting of a small number of parameters determining syntactic structure. Syntactic theory is concerned then mainly with searching for abstract ways of describing languages, in order that underlying similarities can be observed, and parametric differences identified.

However, this view of the core of languages as consisting of rigid innate and universal structures does not seem in accord with much of the available evidence. Languages change gradually over time, children learn languages gradually over a period of several years, and there is a wide range of attested structures in the world's languages, so it seems that learning may be a much more important component in language acquisition than most linguists assume.

The key argument for universal grammar is that children do not receive enough information about the structure of language in order to determine its form. Gold (1967) proved that languages are not 'learnable in the limit' when the learner only has access to positive examples of that language, unless the range of

possible grammars which the learner considers are greatly restricted by some system such as universal grammar.

Pinker (1989) has illustrated one aspect of this learnability problem with a discussion of verb argument structures. He argues that a child who hears constructions such as (1a) and (1b) would then form a rule such that when they observed (1c) they would incorrectly assume that (1d) was grammatical. As children do not generally receive explicit correction as to which constructions are not grammatical they would have no way of recovering from this kind of error.

(1a) John gave a dish to Sam.
(1b) John gave Sam a dish.
(1c) John donated a painting to the museum.
(1d) *John donated the museum a painting.

However Dowman (1998) created a system which was able to learn simple Context Free Phrase Structure Grammars for subsets of a number of languages, despite such arguments as to their learnability. If grammars are assumed to be fundamentally statistical, then it is possible to make inferences about which constructions are unlikely to be absent simply due to chance, and so which postulated constructions are probably incorrect. However, simply using statistical grammars is not enough, as a child learning a language must know at what level to make generalizations, as it is possible to create any number of ad hoc grammars to describe any corpus of data. This problem was solved by using Minimum Coding Length as a metric of the desirability of alternative grammars. (Minimum Coding Length (Ellison, 1992) is similar to Minimum Message Length, and consists of finding the shortest encoding of a grammar specified in terms of the individual symbols which it contains (in this case syntactic categories and words), and then the data specified in terms of the grammar.) Simply applying expectation maximization ultimately produces a grammar which allows the observed sentences and no others as grammatical, as long as the system has no limit on the complexity of grammars.

Programs such as this, and similar grammatical inference systems, such as Stolcke (1994), demonstrate that languages containing many of the features of natural languages, such as gender agreement, verb subcategorizaton, and infinite recursion, can be learned without a universal grammar. Instead the systems need only a learning bias as to the general form which grammars are expected to take, which in the case of Dowman (1998) consisted of a requirement that languages be described using binary branching and non branching phrase structure rules. In the light of this evidence, Gold's concept of learnability in the limit, which requires that a learner be certain of the correct grammar, does not seems interesting from a psychological point of view, where probabilistic inferences may be more useful and robust.

However, the key implication of this research is that it enables us to take a very different perspective on syntactic theory. Instead of always searching for more abstract grammars in order to find underlying universals, it is now possible simply to learn rules specifying the structures present in given languages. It may be that the best syntactic theory is that which assumes the least level of abstraction necessary to account for productivity, seeing as it is more plausible that less abstract grammars could be produced and comprehended at the high speeds necessary during spoken conversation. Hurford (1987) has argued that not all regularities in language need be explained within an ontogenetic account. Some constructions may result from historical processes, and their internal structure need not be analyzed by individual speakers of a language, something which is easy to incorporate into a Minimum Encoding Inference Model of Language, but which causes problems for Chomskyan theories.

By adopting a minimum encoding inference approach to syntactic theory it is possible to gain a new perspective on the problem of explaining acquisition, and one which radically changes our criteria for determining what form a syntactic theory should take. Whilst most work on the machine learning of language has been aimed at producing language technology systems, it seems that machine learning methodologies are essential to making real progress in syntactic theory. While so far analyses of natural language have concentrated on fairly restricted aspects of structure, there is a big potential for applying machine learning techniques to all areas of syntactic theory. Syntactic theories which more closely model the language knowledge of individuals can be applied in explaining social variation, historical change and in clinical linguistics, as well as providing a basis for natural language processing systems.

References

Chomsky, N. (1986). Knowledge of Language, Its Nature, Origin and Use. New York: Praeger.
Dowman, M. (1998). A Cross-linguistic Computational Investigation of the Learnability of Syntactic, Morpho-syntactic, and Phonological Structure. Research Report, University of Edinburgh, Center for Cognitive Science.
Ellison, T. M. (1992). The Machine Learning of Phonological Structure. Doctor of Philosophy Thesis, University of Western Australia.
Gold, E. M. (1967). Language Identification in the Limit. Information and Control, 10:447-474.
Hurford, J. (1987). Language and Number The Emergence of a Cognitive System. Oxford: Basil Blackwell.
Pinker, S. (1989). Learnability and Cognition The Acquisition of Argument Structure. Cambridge, Massachusetts: MIT Press.
Stolcke, A. (1994). Bayesian Learning of Probabilistic Language Models. PhD dissertation, University of California Berkeley.

Strong typing versus declarative bias in ILP

Claude Sammut (University of NSW)

Machine Learning is often viewed as the design of algorithms for searching the space of sentences in a concept description language. As the expressiveness of the language increases, so too does the search space. For this reason, algorithms for learning first-order descriptions must be tightly constrained by some background knowledge. This is usually referred to as a "bias". One method suggested for describing this bias is through the use of strong typing as used in modern functional languages. However, several schemes for providing "declarative bias" have been devised for the more common Horn clause representations used in ILP. This paper will compare the two approaches and show that they are equivalent.

Learning safe programs

Eric Martin (University of New South Wales)

Systems designed to learn logic programs deal with only restricted classes of programs. And the theoretical paradigms that investigate the learnability of logic programs also have to impose stringent conditions on them. When these conditions have a simple expression, the associate class of logic programs is very limited. This is the case, in particular, when local variables are ruled out. Most often, the conditions that determine the class of programs under investigation are very involved, and the resulting class has no topological characterization; its interest will be justified by a mere listing of some of the classical examples that they happen to contain. We will define the class of safe programs, on the basis of a very simple idea: from a syntactic analysis of a program P, it may be possible to prove that for some n-ary predicate R and $1 \leq i, j \leq n$, if the closed atom $R(t_1 \dots t_n)$ is generated by P, then t_i is necessarily "bounded" by t_j (which can mean, for instance, that the size of t_i is at most equal to the size of t_j). Using this knowledge, it may then be possible to prove that for all clauses $C_0 \leftarrow C_1 \dots C_k$ in P and closed atoms $A_0 \dots A_k$, if $A_0 \leftarrow A_1 \dots A_k$ is an instance of $C_0 \leftarrow C_1 \dots C_k$ and if $A_1 \dots A_k$ are generated by P (which implies that A_0 is also generated by P), then $A_1 \dots A_k$ are all "bounded" by A_0. It turns out that the class of safe programs enjoys the property that every recursively enumerable relation is the projection of a relation represented by a safe program. Hence the class of safe programs is both natural and extensive. Moreover, the class of safe programs with a bounded number of clauses and a bounded number of local variables is learnable in the limit from positive data only. We will also explain how we could use a syntactic analysis of the data for an implementation designed to deal with safe programs.

Classification of individuals with complex structure

John Lloyd (Australian National University)

The attribute-value language has been widely used to represent individuals in many practical classification tasks. However, there are also important application domains where this language is not flexible enough. For example, in tasks involving classification of molecules, it is often more convenient to represent

a molecule as a graph. Similarly, in multiple-instance problems, individuals are represented by sets. In this talk, I will discuss the use of higher-order logic for representing individuals with complex structure and how one goes about searching the space of conditions on such individuals.

Use of ordinals to model mistake bounds in learnability of logic programs

Arun Sharma (UNSW) and Sanjay Jain

This talk will discuss the use of constructive ordinals as mistake bounds in the on-line learning model. This approach elegantly generalizes the applicability of the on-line mistake bound model to learnability analysis of very expressive concept classes like minimal models of logic programs. The main result shows that the topological property of effective finite bounded thickness is a sufficient condition for on-line learnability with a certain ordinal mistake bound. An interesting characterization of the on-line learning model is shown in terms of the identification in the limit framework. It is established that the classes of languages learnable in the on-line model with a mistake bound of α are exactly the same as the classes of languages learnable in the limit from both positive and negative data by a Popperian, consistent learner with a mind change bound of α. This result nicely builds a bridge between the two models.

An inference process for identifying structure in text streams

Jon Patrick and Hong Liang Qiao (University of Sydney)

This project aims to build a inference engine that can be trained to identify the structure of a stream of unseen text data. The system will process character data as a stream of records attempting to identify unknown structures. The aim is to focus on text that is known to have some structure in which the boundaries of the structures may be marked but the relationships between structures is unknown. Users should be able to prime the system with as much knowledge as they have in advance to direct the identification of boundaries and the relationships between the structures.

A typical example is the inference of the structure of dictionary entries. In this case the system will need to be able to cope with erroneous data, missing data and irregularly formatted data and intelligently prompt a user to intervene in the inference process as well as allow and record irregular structures. The basic program could be the underlying processor for a series of software functions that are needed by lexicographers. For example the organisation Macquarie Online who publish the Macquarie dictionary have a requirement for 1. A function to automatically tag the entries for subject code categories aligned with topics studied in the high school curriculum; 2. A function to determine the grammatical structures and collocations of words and other linguistic phenomena.

As the system reads records from a dictionary it will need to identify the boundaries of structures from typographical characteristics, especially changes in typography, tag them appropriately for predefined features and then model the structure by building a PFSA. As the construction proceeds or as a post hoc process it will have to infer the dominant structures of the data and report the deviant structures for verification and rectification. The task can be generalised to data of more elaborate structure and for alternative models of the data.

The semantic interpretation of discourse (SID)

Stephen Anthony and Jon Patrick (University of Sydney)

We present a system architecture that deals with the computational linguistic analysis of discourse. The proposed test bed being the analysis of psychotherapeutic transcripts before, during, and after therapy in the hope of determining the effectiveness of the therapy itself using a Neurolinguistic Programming theory of language known as the Metamodel (Bandler & Grinder). Both supervised and unsupervised incremental learning shall be used in conjunction to identify and classify linguistic phenomena contained within turn-taking text in order to identify lexicogrammatical structures deemed to be significant by the Metamodel. Lexicogrammar (Halliday) being both grammar and vocabulary. The document is then marked up using Extensible Markup Language (XML) according to a predefined semantic tag set. The marked up text is then mined for systematic structure in order to produce newly inferred lexicogrammatical rules.

The system has a modular design so as to facilitate plug and play type operation whereby modules

may be interchanged. This modularity allows for instance, the comparison of various machine learning methods by simply initiating the corresponding modules. The key idea behind modularisation is not only the evaluation of learning techniques best suited to processing natural language but also to provide a flexible non-restricted environment which allows users to transfer between application domains and compare results using assorted corpora and language theories.

The project plans to have a number of learning modules available for use. At present the realisation includes transformation-based error driven learning (Brill), decision trees (Quinlan, Cohen), minimum message length, and possibly connectionist methods such as neural networks.

Bandler, R. & Grinder, J. 1975. The Structure of Magic. Science and Behaviour Books: Palo Alto.

Brill, E. 1993a. Automatic Grammar Induction and Parsing Free Text: A Transforma-tion-based Approach. In Proceedings of the 31st Meeting of the Association of Computational Linguistics.

Cohen, W.W. 1995. Fast Effective Rule Induction. In Machine Learning: Proceedings of the Twelfth International Conference: Lake Taho, CA.

Halliday, M.A.K. 1994. An Introduction to Functional Grammar Second Edition. Edward Arnold: Great Britain.

Quinlan, J. 1992. C4.5: Programs for Machine Learning. Morgan Kaufmann: San Mateo, CA.

Decision trees, margins and Brownian motion

Yoav Freund (AT&T Labs — Research)

Adaboost is a learning algorithm that, in recent years, has been gaining a name as one of the best off-the-shelf learning algorithms. In fact, Adaboost is not exactly a learning algorithm, rather, it is a method for improving or "boosting" the performance of a given learning algorithm, often called the "base learner". The most popular base learners so far have been algorithms for learning decision trees, such as CART and C4.5. In the first part of the talk I will describe a new learning algorithm which integrates a tree-learning algorithm with a boosting algorithm. This algorithm generates a new type of classifier, which we call "alternating trees" and is a significantly more powerful representation than decision trees or boosted decision trees.

One of the surprising phenomena associated with boosting is that it is much more resistant to overfitting than expected. In recent work we have demonstrated that this behavior can be explained using the notion of "margins". In fact, looking at adaboost from this perspective it becomes apparent that adaboost is performing a type of gradient descent with respect to a potential function that is exponential in the margin. It also highlights the problem that the exponential function is a poor approximation to the actual target function, which is the classification error step function. In the second part of the talk I will describe a new boosting algorithm which is based on the equation for the time evolution of Brownian motion. This algorithm approximates the classification step function with an error function (erf) rather than an exponential.

Sparse kernel feature analysis

Alex J. Smola (Australian National University), Olvi L. Mangasarian (University of Wisconsin), and Bernhard Schölkopf (Microsoft Research)

Kernel Principal Component Analysis (KPCA) has proven to be a versatile tool for unsupervised learning, however at a high computational cost due to the dense expansions in terms of kernel functions. We overcome this problem by proposing a new class of feature extractors employing ℓ_1 norms in coefficient space instead of the reproducing kernel Hilbert space in which KPCA was originally formulated in. Moreover, the modified setting allows us to efficiently extract features maximizing criteria other than the variance much in a projection pursuit fashion.

Maximum margin learning vector quantization

Lawrence Buckingham and Shlomo Geva (Queensland University of Technology)

In this paper we show that there is a close connection between Support Vector Machines (SVM) with gaussian kernels [Vapnik, 1995] and Learning Vector Quantisation (LVQ) [Kohonen, 1988]. To do this, we

construct a classifier similar to a SVM, which has LVQ as a limiting case. We see that the LVQ training algorithms can be derived by maximising the margin of this classifier, noting that the main qualitative difference between LVQ and SVM is that SVM training holds the centre of each kernel fixed and adjusts mixing weights, while LVQ permits kernel centres to move but holds the mixing weights constant.

Kohonen's LVQ algorithms are a closely related set of methods for training a compact nearest neighbour classifier from examples. The primary LVQ algorithm (called LVQ1) is a very efficient procedure that forms a non-parametric model of the distribution of the training data, while the LVQ2.1 and LVQ3 variants can be used to further refine a classifier produced by LVQ1 training to increase classification accuracy. LVQ1 can be derived by an argument based on average distance from training points to prototypes, but the reasoning behind LVQ2.1 and LVQ3 modifications is largely heuristic. Moreover, LVQ2.1 and LVQ3 require the user to furnish additional parameters to control learning.

In the present work we picture the LVQ classifier as a normalised mixture of gaussian kernels, controlled by a single shared kernel width parameter σ, which in the limit of small σ becomes a nearest neighbour classifier. We derive a training algorithm for this classifier which maximises the average margin over the training set, and illustrate the circumstances under which each of the three LVQ variants can be recovered from the new algorithm. The new algorithm unifies the three flavours of LVQ, reducing the required number of parameters that need to be supplied by the user and capturing the best features of all three algorithms.

Reinforcement learning in continuous action space: a bicephal solution

Frederic Maire (Queensland University of Technology)

The classical Reinforcement Learning (RL) algorithms require a discrete and finite action space. From a programming point of view, in RL, a discrete action space is more convenient than a continuous action space; the determination of a maximizing action with respect to the action-value can then be done by simply looking-up a table. However, in many applications, the action space of the agent is continuous. A simple solution consists in discretizing the action space. But, the discretization itself poses some problems (loss of precision in the actions, determination of an adequate discretization, etc...). We propose a neural solution using a pair of coupled networks. The first network implements the policy of the agent, the second network implements the action-value function. We have adapted Q-learning to this neural system. The key element of the adapted Q-learning algorithm is a gradient ascent on a subset of the entries (corresponding to the action) of the second network to improve the current policy. We will present simulation experiments on a ball chasing problem.

Direct gradient-based reinforcement learning

Jonathan Baxter and Peter Bartlett (Australian National University)

Many control, scheduling, planning and game-playing tasks can be formulated as reinforcement learning problems, in which an agent chooses actions to take in some environment, aiming to maximize a reward function.

Formally, reinforcement learning problems can be modelled as Markov Decision Processes (MDP's). For small enough state spaces, the techniques of dynamic programming enable an optimal policy to be found by solving Bellman's equations for the value function, and then using the value function to generate a policy by choosing in each state the action leading to the state with the highest expected value.

However, for many problems the state space is far too large to find the optimal value function, hence the emphasis in reinforcement learning has been on finding approximatitions to the value function within a restricted class such as neural networks. While there have been a number of empirical successes with this approach, it suffers from a lack of fundamental theoretical gurantees on the performance of the policy generated by the approximate value function.

This talk describes an alternative approach to reinforcement learning, which we call direct reinforcement learning. We consider policies that are defined by parameters – they might define approximate value functions that are used to generate a policy by some form of look-ahead, or they might directly parameterize the agent's policy. Rather than attempting to find accurate value estimates and then use these to generate

a policy, we instead directly adjust the parameters to improve the average reward. Specifically, we compute the gradient of the average reward with respect to the parameters, and then use gradient ascent to generate a new set of parameters with increased average reward.

The talk will present an algorithm for computing accurate approximations to the gradient of the average reward from a single sample path. We use these estimates in a conjugate-gradient ascent algorithm that uses a novel line-search routine, relying solely on gradient estimates. In a variety of domains, including a communications network call admission problem and a simple dynamic system control problem, this algorithm rapidly finds optimal or near-optimal solutions.

Reinforcement learning in networks of spiking neurons

Peter Bartlett and Jonathan Baxter (Australian National University)

In reinforcement learning problems, an agent attempts to learn appropriate actions so as to maximize a reward signal. We present an algorithm for reinforcement learning in networks of spiking neurons. This algorithm is qualitatively similar to Hebb's postulate in the way it modifies synaptic connection strengths . It requires only simple computations (such as addition and leaky integration), involves only a single global reward signal together with quantities that are available in the vicinity of the synapse, and leads to synaptic connection strengths that give locally optimal values of the long term average reward. The reinforcement learning paradigm is sufficiently broad to encompass a variety of learning problems. Simulations illustrate that the approach is effective for simple pattern classification and motor learning tasks.

Computational models of the basal ganglia and cerebellum for sensorimotor control

Marwan Jabri (University of Sydney)

The goal of this study is to develop a sensorimotor system based on computational models of a cerebellum and basal ganglia operating on a micro robot. The experiment requires the robot to navigate on a flat surface and to track a target moving in a predictable manner by using its camera. Learning of motor control utilizes the predictive Hebbian reinforcement-learning algorithm in the basal ganglia module. Learning of sensory predictions makes use of a combination of long-term depression (LTD) and long-term potentiation (LTP) adaptation rules within the cerebellum module. The basal ganglia module uses the visual inputs to develop sensorimotor mapping for motor control, while the cerebellum module utilizes robot orientation and target spatial inputs to predict the location of the moving object. We propose several hypotheses about the functional role of cell populations in the cerebellum and argue that mossy fiber projections to the DCN play a coordinate transformation role and act as gain fields. We propose such field is learned early in the brain development with respect to the activity of the climbing fiber. Proprioreceptor mossy fibers projecting to the DCN and providing robot orientation with respect to a reference system are involved in this case. Other mossy fibers carrying visual sensory input provide visual patterns to the granule cells to construct a binary basis representation. The combined activities of the granule and the Purkinje cells store spatial representations of the target patterns. The combinations of mossy and Purkinje projections to the DCN provide a prediction of the location of the moving target taking into consideration the robot orientation. The cerebellum module learns to predict future target positions in spatial coordinates and transforms it into patterns on the visual field. The overall system exhibits a behavior similar to that of primates in performing anticipatory tasks. Results of lesion simulations based on our model show degradations similar to those reported in cerebellar lesion studies on monkeys.

Selective attention adaptive resonance theory

Peter Lozo (Defence Science and Technology Organisation)

In this presentation I will review some neurophsyiological data on selective visual attention and memory guided search and then present a real-time neural network theory and model of visual pattern recognition that was developed to (i) address the problem of object recognition in cluttered visual and IR images and (ii) to explain the role of the massive feedback pathways in the visual cortex. Building onto the Adaptive Resonance Theory (ART) of Stephen Grossberg and the ART based neural networks of G. Carpenter and S.

Grossberg, Selective Attention Adaptive Resonance Theory (SAART) proposes novel top-down feedback pathways (top-down presynaptic facilitation) that regulate the bottom-up synaptic signal transmission gains.These new interactions enable object recognition in cluttered images. A more advanced version of the network (Advanced SAART) deals with the problems of memory guided search and translation, size and orientation invariant object recognition.

Clustering to enhance FSA extraction from recurrent networks

Dylan Muir, M. Towsey and Joachim Diederich (Queensland University of Technology)

When training an Artificial Neural Network, we want to gain some understanding of the network's representation of the training problem. A common way to represent the knowledge of a Recurrent Neural Network (RNN) is as a Finite State Automaton (FSA). FSA extraction from RNNs is usually done after successful training; states are inferred from clustering of the hidden layer activations. These FSAs are a representation of rules identified by the neural network from the training set; however, the states (areas in hidden layer activation space) are rarely clean, and often overlap. This problem can be addressed by introducing more states, however this reduces the comprehensibility of the FSA.

If it is possible to represent the training data as an FSA, we would like the RNN to learn as simple an FSA as possible. Previous work has induced FSAs representing the simple Tomita languages (Das & Mozer, 1999). Here we investigate the generalisation of these principles to a more complex (natural spoken language) dataset, which cannot be represented entirely as a small deterministic FSA. Our goal is to train networks representing FSAs with as few states as possible, without sacrificing prediction accuracy.

We utilise a method of enforcing clean FSA extraction by quantising the hidden layer activations. Analysis is done on a one-step-lookahead task from a large spoken language corpus. The network is a variant of the standard Elman SRN (Simple Recurrent Network) architecture trained with back-propagation, except with hidden layer activations being forced towards their closest cluster at every pattern presentation. Re-clustering is done on a set of hidden layer activations gathered over a few epochs, using an adaptive clustering algorithm. The codebook generated by this process is used for subsequent training until the next re-clustering period.

This same clustering algorithm is used to extract an FSA from a standard Elman network trained without on-line clustering. A comparison between these two FSAs reveals a similar prediction score with a greatly reduced number of states.

References: Das, S., Mozer, M.: "Dynamic On-Line Clustering and State Extraction: An Approach to Symbolic Learning", Neural Networks, Vol 11, pp 53-64, 1998

Data sharpening

Peter Hall (Australian National University)

Methods are suggested for improving the performance of a range of statistical methods by adjusting the data, rather than adjusting the estimator. A major potential of this approach lies in its ability to enhance performance in relatively complex settings, for example when the number of dimensions is high. There, refinement of the estimator can be difficult, but often the data can be "tweaked" in a rather simple way and analysed using a particularly simple method, to achieve much the same affect as employing a more complex method.

Smoothing and approximation techniques applied to data mining problems

Stephen Roberts (Australian National University)

In data mining it is often necessary to provide an efficient prediction of a response variable, where there may be many predictor variables (tens to hundreds) and there are megabytes of data (or more). Many other techniques can be based on good predictive modelling, for instance classification, density estimation and clustering. The aim of the algorithm developer is to produce a scalable algorithm, that is an algorithm which scales linearly with the number of data items, and can deal with the large dimensionality of the data. In this talk we will describe how we are using finite element methods, multigrid methods, additive models, sparse grids and parallel programming techniques to develop a scalable multivariate smoothing algorithm

which can be used for predictive modelling of typical data mining datasets.

The effects of class size on template matching

Hari Koesmarno and Warwick Graco (Health Insurance Commission)

This research demonstrates the effects of both unsupervised learning and supervised learning in identifying patients who are doctor shoppers in patient data. Doctor shoppers are patients who have dependencies on prescription drugs and visit many doctors in different geographic locations to obtain the prescriptions they require. The unsupervised learning involved clustering a training set consisting of patient profiles and the profile of best fit for each cluster (also called a template) was risk classified by an expert in terms of indicating the risk the patient was a doctor shopper. The supervised learning used the results of the unsupervised learning where the templates were matched with patient profiles in a test set and the profiles were given the classification of the closest matching template. The profiles in the test set were also risk-classified by two other experts. Both experts worked as a team to give composite classifications. The classifications of the templates were compared with the composite classifications of the two experts to determine the accuracy of the template matching. Research was conducted into the effects template class sizes have on the accuracy of the classifications of profiles in the test set. Class size refers to the number of templates in a risk classification or class. The results showed that class size affects the accuracy of the classifications and that large, even-sized classes appears to be the best option with template matching. The results and implications of the research are reported.