DEPARTMENT OF COMPUTER SCIENCE
Royal Holloway, University of London

Postgraduate Colloquium 2013
Wednesday 5 June

Queens Building Lecture Theatre

Session 1: Machine Learning I

Valentina Fedorova
Conformal prediction under hypergraphical models

Conformal predictors are usually defined and studied under the exchangeability assumption. However, their definition can be extended to a wide class of statistical models, called on-line compression models, while retaining their property of automatic validity. The talk is about conformal prediction under hypergraphical models. In the talk I define the hypergraphical models and introduce a family of conformity measures for the models. The measures are used to construct special conformal predictors. Empirical results for the conformal predictors will be reported showing their performance on benchmark LED data sets. The experiments confirm that conformal predictions under hypergraphical models are valid and show that they are more efficient as compared to conformal predictions under the exchangeability assumption.

Meng Yang
Utilize Additional Information by Conformal Predictors

In many supervised learning applications, the existence of additional information in training data is very common. However, traditional learning method cannot utilize this kind of information because additional information are not available in test set. Vapnik introduced a new paradigm LUPI which provides the learning paradigm with additional information and take advantage of it through SVM+ method in off-line mode. Following his idea, we would like to use additional information by conformal predictor. Conformal predictor is recently developed learning framework which allows to make prediction with reliable measures of confidence and its predictions have validity property. We hope to extend it to maintain the validity property and take advantage of additional information in both on-line and off-line mode. The method will be applied for medical diagnosis.

Chenzhe Zhou
SVM Venn Machine with k-Means Clustering

In this presentation, I will introduce a new method of Venn Machine taxonomy design based on support vector machines and k-means clustering for both multi-class and binary cases. Then I will compare this algorithm to some other multi-probability predictors including our previous Venn Machine and a recently developed algorithm. The algorithms were tested on a range of real-world data sets.

Antonis Lambrou
Reliable probabilistic outputs for large datasets

Venn Predictors (VPs) are machine learning algorithms that can provide well calibrated multiprobability outputs for their predictions. An important drawback of Venn Predictors is their computational inefficiency, especially in the case of large datasets. In this work, we investigate and propose the use of Inductive Venn Predictors (IVPs), which can overcome the computational inefficiency problem of the original Transductive Venn Prediction framework. We develop an IVP based on the Sequential Minimal Optimisation (SMO) algorithm and perform a detailed comparison of its time efficiency, accuracy, and quality of probabilistic outputs with those of the original SMO with Logistic Regression and of the corresponding Transductive Venn Predictor (TVP). The results demonstrate that our method provides well calibrated results while maintaining high accuracy. The IVP outperforms the original SMO and TVP methods in terms of time efficiency, while also providing well-calibrated probabilistic estimates, in contrast to those of the original SMO method. Another observation is that the probability intervals of the IVP are tighter than those of the TVP.

Session 2: Machine Learning II

Jiaxin Kou
High Dimensional Visualization on Support Vector Machine Research

By Data Altas approach enhanced by Proxigram, we could take advantage of our visual system to discover and explore high dimensional data space for inspiration as well as insight. This year we applied this method on SVM machine learning problem, visualising its data distribution, support vector, prediction error pattens and spotting underfitting and overfitting area, which leads to some interesting data stories.

Tim Scarfe
Merging Time Series with Specialist Experts

The paper describes an application of specialist experts techniques to prediction with side information. We pick vicinities in the side information domain to create elementary time series, use standard prediction techniques to predict for those elementary series, and then merge the predictions using specialist experts methods. Prediction with expert advice bounds ensure that optimal vicinities are selected dynamically. The algorithm is tested on the problem of predicting implied volatility of options and proves to be a viable alternative to on-line regression.

Ulrich Schaechtle
Multi-dimensional Causal Discovery

We propose a method for learning causal relations within high-dimensional tensor data as they are typically recorded in non-experimental databases. The method allows the simultaneous inclusion of numerous dimensions within the data analysis such as samples, time and domain variables construed as tensors. In such tensor data we exploit and integrate non-Gaussian models and tensor analytic algorithms in a novel way. We prove that we can determine simple causal relations independently of how complex the dimensionality of the data is. We rely on a statistical decomposition that flattens higher-dimensional data tensors into matrices. This decomposition preserves the causal information and is therefore suitable for structure learning of causal graphical models, where a causal relation can be generalised beyond dimension, for example, over all time points. Related methods either focus on a set of samples for instantaneous effects or look at one sample for effects at certain time points. We evaluate the resulting algorithm and discuss its performance both with synthetic and real-world data.

Session 3: Agent Technology

Paulo Ricca Goncalves
Open Objects Framework

We live in a world that is becoming increasingly populated by computationally-capable objects (such as your college card or your mobile phone) connected in various ways between them (such as wifi or bluetooth). One question that often arises is "can we connect these objects in ways that they can cooperate and help us with what we do in a our daily life?" We present the Open Object framework, a lightweight decentralised model for orchestrating the collective behaviour of physical objects with computational capabilities.

Ataul Munim
Introducing the concept of Infrastructure Agents for the OpenGOLEM platform

The OpenGOLEM project aims to facilitate the creation of multi-agent applications through an array of decoupled components, helping developers and researchers implement their designs with minimal effort. As an extension to this initiative, we introduce the concept of infrastructure agents as a means of further compartmentalising of roles and components within a given system.

Bedour Al-Rayes
An Agent Architecture for Adaptive Decision-Making in Negotiation Environments

Popular e-commerce market places are normally based on fixed prices (e.g. Amazon books) or mediated via negotiation mechanisms based on auctions (e.g. e-Bay). However, for a large class of goods the fixed price model is not flexible enough since one cannot negotiate the price. Similarly, the limitations of auctions for certain goods are well known: they are time consuming, communication is unidirectional and participants cannot adapt their strategies to the behaviour of their negotiation opponents. To address these limitations, we advocate the deployment of software agents that are capable of bargaining to allocate goods faster by communicating in a bidirectional way, through offers and counter offers. To the best of our knowledge, there are no existing bargaining agents available on the Internet. Consequently, this suggests the need to conduct more research in this area and develop bargaining agents that negotiate in multiple electronic markets on behalf of their users. We contribute to the state of the art by proposing a novel agent-architecture that represents a symbolic decision making component to support the deployment of concurrent negotiation. The architecture is a revised version of previous work with the KGP model that incorporates the environment, opponents and self models. Our work focuses on the specification of a domain-independent decision-making capability that can be combined with a new concurrent protocol as an extension to the well-known alternating offers protocol. The skeleton of the decision-making capability described illustrates how to link it to the agent strategies, utilities and preferences using a

Prolog-like meta-program. The work prepares the ground for supporting decision-making in multiple concurrent negotiations that is more lightweight than previous work and contributes towards a fully developed model of the architecture, to be presented in the future.

Ionut Tutu
The Logic Programming of Service-Oriented Computing

Service-oriented computing is a recent paradigm focusing on computation in data processing infrastructures that are globally available, and in which software applications can discover and bind dynamically to services offered by providers. The object of this talk is to describe how aspects specific to the logic programming paradigm can be used to capture the declarative and operational semantics of service-oriented computing. To this purpose, we develop an integrated algebraic framework that constitutes the primary factor in defining logical specifications of services and models of those specifications (corresponding to orchestrations of components that depend on externally provided services), as well as in explaining how the discovery of services and the binding of their orchestrations to client applications can be expressed in logical terms. This allows us to interpret the discovery of a service that can be bound to a client application as the search for a clause that can be used in computing an answer to a query, and to describe the process of binding services and the reconfiguration of applications as service-oriented counterparts of unification and resolution.

Session 4: Languages; Discrete Optimisation Algorithms; Bioinformatics

Robert Walsh
Abstracting the C# Grammar

C# is a language that was initially released by Microsoft in 2002. Both the ISO/IEC and ECMA language specifications for C# formally describe the structure and syntax of C# and is accompanied by a corresponding 'concrete syntax grammar' (CSG). However the description of the semantics is less precise and lacks formalism, and this is a situation present in many language specifications.
The aim of the PLanCompS project is to provide techniques and tooling which allow all designers of both general purpose and domain specific languages to use formal semantics descriptions by providing reusable semantics fragments which may be deployed by non-specialists. To support the formal semantics specification, the CSG of the language needs to be rewritten into a more structurally concise 'abstract syntax grammar' (ASG). Using C# as a case study, this presentation compares the CSG given in the presentation and, alongside, a hand-constructed ASG being used for the semantic specification.
From this, we identify common patterns and thus build up a collection of generalisable transformations which can be used to transform the former grammar into the latter.

Gabriele Muciaccia
Polynomial Kernels for λ-extendible Properties Parameterized above the Poljak-Turzik Bound

Poljak and Turzík (*Discrete Mathematics* 1986) introduced the notion of λ-extendible properties of graphs as a generalization of the property of being bipartite. They showed that for any $0<\lambda<1$ and λ-extendible property Π, any connected graph G on n vertices and m edges contains a spanning subgraph $H\in\Pi$ with at least edges. The property of being bipartite is λ-extendible for $\lambda=1/2$, and so the Poljak and Turzík bound generalizes the well-known Edwards-Erdős bound for Max-Cut. Other examples of λ-extendible properties include: being an acyclic oriented graph, a balanced signed graph, or a q-colorable graph for some $q\in N$.

Mnich et. al. (*FSTTCS* 2012) defined the closely related notion of *strong* λ-extendibility. They showed that the problem of finding a subgraph satisfying a given strongly λ-extendible property Π is fixed-parameter tractable (FPT) when parameterized above the Poljak-Turzík bound—*does there exist a spanning subgraph H of a connected graph G such that H∈Π and H has at least edges?* —subject to the condition that the problem is FPT on a certain simple class of graphs called *almost-forests of cliques*. This generalized an earlier result of Crowston et al. (*ICALP* 2012) for Max-Cut, to all strongly λ-extendible properties which satisfy the additional criterion.

We settle the kernelization complexity of nearly all problems parameterized above Poljak-Turzík bounds, in the affirmative. We show that these problems admit quadratic kernels (cubic when $\lambda=1/2$), *without using* the assumption that the problem is FPT on almost-forests of cliques. Thus our results not only remove the technical condition of being FPT on almost-forests of cliques from previous results, but also unify and extend previously known kernelization results in this direction. Our results add to the select list of *generic* kernelization results known in the literature.

Horacio Caniza Vierci

A disease similarity measure for Mendelian diseases in Man

An important problem in medicine is determining the similarity between two diseases. Based on the premise that similar diseases share a similar molecular background, determining which diseases are similar to one another might provide insight into them, including the transfer of knowledge from well- to less- studied diseases.

The Online Mendelian Inheritance in Man (OMIM) is a high quality database of phenotype descriptions, their molecular basis and supporting publications for human Mendelian diseases. Developed as a catalogue for medical professionals, its free-text description of phenotypes hinders automatic data extraction.

We have mined OMIM and extracted each entry s related publication, which we later annotated with the Medical Subject Headings (MeSH) vocabulary. Built as a comprehensive taxonomy, MeSH indexes journals and books in the life sciences. We have performed semantic similarity calculations on the MeSH-annotated publications retrieved from OMIM and have obtained a proxy similarity measure between diseases. Preliminary results indicate that the proposed measure effectively quantifies disease relatedness.