ROYAL HOLLOWAY UNIVERSITY OF LONDON

CSPRC 2014

Abstracts

14th Computer Science Postgraduate Research Colloquium

Ionuţ Ţuţu Jamie Al-Nasir

Royal Holloway University of London
Department of Computer Science

4th June 2014

Contents

Session 1: Machine Learning

Session 2: Logic, Types and Language Engineering

Session 3: Bioinformatics

Session 4: Intelligent Systems

Poster Presentations

Preface

Welcome to the 14th Computer Science Postgraduate Research Colloquium (2014). This colloquium serves as a forum for interaction between the faculty, staff and students engaged in the various different disciplines of research in computer science at Royal Holloway.

This year's event is the result of a staff-student collaboration, and has been coordinated for the first time by the students. The work presented showcases the innovative ideas, diversity and excellence of postgraduate research that is conducted here, within the Department of Computer Science at Royal Holloway University of London.

For this year, we have organized talks and posters into four themes:

- Machine Learning,
- Logic, Types and Language Engineering,
- Bioinformatics, and
- Intelligent Systems.

We hope you enjoy the day, and we thank you for taking part.

30th May 2014
Royal Holloway

Ionuț Țuțu
Jamie Al-Nasir

Probabilistic prediction under unconstrained randomness

Chenzhe Zhou

When we learn from experience, we always want that there is some stability in the environment in order for there to be something to learn in that environment. To assume that all the examples are generated randomly from some fixed probability distribution on a fixed space of possible examples is the traditional way of making the idea of a stable environment precise. And we say all the examples are independent and identically distributed (i.e. the i.i.d. assumption). Furthermore, if we only know the space of examples and all the examples are under i.i.d. assumption when we implement a learning task, we say that we are learning under unconstrained randomness. Although it is impossible to estimate the true conditional probabilities under the conditions stated, we could still give probabilities that have a frequentist justification. In this presentation, I will mainly introduce a multi-probabilistic predictor under unconstrained randomness called Venn predictor together with some taxonomy designs used in Venn predictor. Some benchmark performances on real world data sets and comparisons with other algorithms will also be given in this talk.

Anomaly detection of trajectories with kernel density estimation by conformal prediction

James Smith

Anomaly detection is a large area of research in machine learning and many interesting techniques have been developed to detect "abnormal" behaviour of objects. We define "anomalies" to be entities in the data that do not conform to a typical behaviour. These non-conforming entities are often called "anomalies" or "abnormalities" or "outliers". Recently conformal anomaly predictors have emerged which allow the detection of the non-conformal behaviour of objects using some measures of non-conformity. This technique also has an advantage in delivering provably valid confidence measures under the exchangeability assumption that is usually weaker than those traditionally used. Conformal predictors produce prediction sets such that the probability of an example belonging to a prediction set is greater than a significance parameter ϵ.

In this presentation, we describe the problem of anomaly detection in the maritime domain, that deals with trajectories of vessels to detect suspicious behaviours for example: a sudden change of direction, or speed, or anchoring, etc. The data used in experiments were obtained from real Automatic Identification System (AIS) broadcasts. Automatic Identification System (AIS) is a system for vessels to broadcast their reported locations from GPS receiver.

It is difficult to acquire labelled examples of anomalous vessel trajectories. One previous approach of studying the performance of non-conformity measures is to create artificial 'anomalies' and use these in a test set alongside previously observed vessels trajectories that are considered 'normal'. The true positive and false positive rates for various ϵ can be used to measure performance.

An alternative is to use the unsupervised setting in which we assume as much of the feature space as possible is anomalous. We propose a new metric ‘average p-value’ for measuring the performance of conformal anomaly detectors in this setting, it is independent of the significance parameter ϵ.

We compare and present the results of both the k-nearest neighbour and kernel density non-conformity measures, through the use of dimensionality reduction with both metrics and settings.

Causal discovery with Gaussian processes

Ulrich Schaechtle

We present a method for causal discovery using Gaussian processes in the context of additive noise models. We introduce a simple Gaussian process that is comprised of a linear term, a homoscedastic Gaussian process component and a heteroscedastic Gaussian process component. These are optimised to model the causal generative process. The interplay of each of these three components allows us to identify cause and effect without making too strong assumptions with regards to the nature of the true underlying causal function and noise. The model also comes with a strong predictive capability.

What is a logic-programming language?

Ionuţ Ţuţu

The logic-programming paradigm is a prime example of how the integration of declarative and operational aspects of logical systems can be used to provide a single unified framework for both specification and programming. In essence, programming in logic amounts to giving appropriate axiomatic formalizations for the computable functions of interest, which can then be executed by means of goal-directed deductions that are performed according to a fixed strategy.

In this talk, we examine common features of various conventional logic-programming languages, ranging from the most traditional variant of the paradigm – defined over the Horn-clause fragment of relational first-order logic – to equational, higher-order, and object-oriented logic programming. Based on these, we propose an abstract model-theoretic framework that allows us to develop and conduct research into logic programming over an arbitrary logical system, that is without concrete models, sentences, satisfaction or deduction, and thus to explore the logic-programming paradigm for other, less conventional formalisms such as the logic of orchestration schemes used in service-oriented computing.

Our study is based on abstractions of notions such as logic program, clause, query, solution and computed answer, which we develop over Goguen and Burstall's theory of institutions. These give rise to a series of concepts that formalize the interplay between the denotational and the operational semantics of logic programming. We investigate properties concerning the satisfaction of quantified sentences, discuss a variant of Herbrand's theorem that is not limited in scope to any construction of logic programs, and define a sound and conditionally complete procedure for computing solutions to queries.

Type-theoretic description of higher categories

Fjodor Part

Inadequacy of set-theoretic foundations of mathematics as a basis of a practical tool for writing proofs in computer checkable form, coupled with increasing demand for such a tool, has drawn attention of mathematicians to the problem of developing new foundations. Recent discoveries of higher categorical models for intensional dependent type theory (ITT) have given birth to Univalent Foundations.

Central idea is that types in ITT can be interpreted as weak ∞-groupoids or, equivalently by Grothendieck's hypothesis, as homotopy types. Such interpretation suggests an idea to augment ITT with the new axiom, namely the Univalence axiom, to make the system more expressive as a language of ∞-groupoids and, ideally, to make it internal logic of ∞-topoi. The resulting system is called Homotopy type theory (HoTT).

Plenty of rather sophisticated constructions and proofs has been formalised in elegant way in HoTT. Nevertheless, there are some higher categorical objects, such as simplicial objects, definition of which faces serious difficulties, such as necessity to specify infinite amount of coherences. This motivated development of other type theories and so far just Homotopy type system (HTS) has been announced as one, capable to describe simplicial objects. In the talk, I will present this system, explain some of it's flaws and suggest ways to build a system without them.

Another line of our research I will be talking about is employing intensional type theories to obtain internal logic of higher categorical structures. That is, obtaining further correspondences, such as between HoTT and ∞-topoi, as well as developing necessary tools, requiring for stating formally and proving, that a type theory provides internal logic for higher categorical structure.

Extending lexical analysis to generalised parsing

Robert Michael Walsh

Traditional parser generators start with an input sequence of characters. This input is processed by a *lexical analyser*, which produces a sequence of symbolic *tokens*. This rewritten sequence is then sent to a *syntax analyser* as input which, given a grammar, produces a structural representation of the sequence (typically a derivation tree or shared packed parse forest).

Typical lexical analysers, such as Lex, return a single sequence of tokens. However, there may exist multiple sequences of tokens (or lexicalisations) that a particular lexer scheme could produce for a given input sequence of characters. To resolve these *lexical ambiguities*, Lex applies the following rules:

1. If a token definition can match more of the input, then choose the match that consumes the most input (longest match).
2. If multiple tokens match the same subsequence of characters, then choose a token according to some partial ordering of the tokens.

However, a longest match of the tokens may not always be desired. In particular, there are situations (that will be demonstrated in this talk) where longest match creates inconsistencies in some widely used programming languages.

Some generalised parser tools, such as SGLR, take a *scannerless parsing* approach where lexical analysis and syntax analysis are merged into a single stage. This removes the need to classify sequences of characters before their context within the parsing grammar is known. However, this greatly increases the size and complexity of the resulting structural representation.

The approach presented in this talk instead generates a graph representing all the possible lexicalisations. A modified parser can then take this graph as input and consider all options presented. As will be seen, this gives the full range of parses available in the scannerless parsing approach whilst still allowing the traditional approaches towards reducing the ambiguities, resulting in a more concise structural representation.

A disease similarity based on an ontological analysis of phenotipic descriptions

Horacio Caniza Vierci

Over 7000 Mendelian diseases are catalogued in OMIM. These diseases range from the very well-known like sickle-cell anaemia to the more obscure ones like the Mseleni joint disease. Obtaining an effective similarity measure between diseases is a very important problem as it would help transfer knowledge from the well-studied diseases to the obscure ones, with the final aim of predicting disease causing genes.

We have mined the OMIM catalogue of inherited diseases and extracted each entry's related publication. These publications were 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.

Our results show a significant improvement over the state of the art. The results also show that our novel method effectively exploits existing phenotipic information to characterise the genotype-phenotype relationship.

Interactive high dimensional data exploration and analysis using Proxigram Viewer

Jiaxin Kou

To leverage limited capabilities of our previous static Proxigram, we built a visual tool called Proxigram Viewer from scratch with Python Language, bringing various interactive features to high dimensional data exploration and analysis. In order to obtain distinctive data stories, recently we visualised a real world dataset – Protein-Protein Interaction Network. In experiments, our application gives a better landscape of yeast protein complexes in 2D space and an alternative way to inspect the protein clustering results produced by ClusterOne algorithm, which could enable many possibilities for future research.

Adaptive decision-making model in a dynamic negotiation environment

Bedour Alrayes

The proliferation of e-markets on the web demands for automated mechanisms of negotiation that can relieve users from following multiple bids over different web sites. One line of research that is being proposed in this setting is to develop negotiation models that support a software agent representing a user in an e-market to negotiate with other agents bilaterally and concurrently.

One of the main challenges in designing such a negotiation model is to deal with the openness and dynamic nature of the environment in terms of the agent architecture, the negotiation protocol used and the strategy that the agent must adopt to maximize its utility (more precisely that of the user it represents).

The main goal of this work is to develop an adaptive decision making model for concurrent bilateral negotiations with emphasis on in terms of the agent architecture, the negotiation protocol used and the strategy that the agent must adopt to maximize the utility of the user it represents. To test the model we also seek to develop an agent negotiation simulator, to test the strategy and compare our work with the current state of the art.

In this talk we will report on two main contributions:

1. CONAN: this is a heuristic agent model for concurrent bilateral negotiations in electronic markets that are open, dynamic and complex. Existing models often omit the factors determining when a market environment is open or how an agent evaluates progress in bilateral negotiations. Such omissions in turn damage the offer-making ability of an agent and consequently the number of successful negotiations that this agent can achieve. Negotiation experiments indicate that CONAN outperforms other agents that

rely on the current state-of-the-art by a significant amount in terms of the utility gained during a negotiation.

2. RECON: is an experimental simulation platform that supports the development of software agents interacting concurrently with other agents in negotiation domains. Unlike existing simulation toolkits that support only imperative negotiation strategies, RECON also supports declarative strategies, for applications where logic-based agents need to explain their negotiation decisions to a user. RECON is built on top of the GOLEM agent platform, specialized with a set of infrastructure agents that can manage an electronic market and extract statistics from the negotiations that take place. We evaluate the performance of RECON by showing how by increasing the number of agents in a simulation affects the agents' time to make an offer during negotiation.

Semi-automatic indoor fingerprinting database crowdsourcing with continuous movements and social contacts

Khuong Nguyen

Indoor localisation helps monitoring the positions of a person inside a building, without GPS coverage. In the past decade, much research effort has been invested into Indoor Fingerprinting, which is considered one of the most effective indoor tracking methods to date. In recent years, some researches started looking at crowdsourcing the fingerprinting database with the contributions from indoor users via mobile phones or laptop PCs. However, the crowdsourcing process was greatly limited due to the lack of indoor reference, in contrast to the widespread use of GPS reference for outdoor crowdsourcing. In this talk, we propose a novel idea to crowdsource the fingerprinting database without any preset infrastructure, landmarks, nor using any advanced sensors. Our idea is based on the observations that the users often carry a mobile phone with them, and there are multiple social contacts amongst those users indoor. First, we exploit the user's continuous movement indoor to refine the location prediction set. Second, we use a unique concept to detect the indoor social contacts with NFC by tapping the back of the 2 phones together. Third, we propose a novel idea to combine this social contact with the user's continuous movements to identify the exact entries in the fingerprinting database that need updating for crowdsourcing.

Investigation into the annotation of sequencing protocol steps on the Illumina platform in genomic data repositories

Jamie Al-Nasir

The work-flow for the production of high-throughput sequencing data from nucleic acid samples is a complex one. There are a series of protocol steps in the preparation of samples for next-generation sequencing. The quantification of the bias due to specific combinations of these protocols remains to be determined. We outline the typical sequencing work-flow on Illumina platforms (HiSeq 1000/2000, HiScanSQ, Genome Analyzer IIx, MiSeq). We discuss potential sources of bias in DNA fractionation, blunting, phosphorylation, adapter ligation and library enrichment and the variation in these protocols in both wet-lab and dry-lab processes. In order to understand possible sources of bias that exists in publicly deposited data sets of this type, it is important that the associated meta-data have information about what specific protocols were used.

We examined the experimental meta-data of the SRA (Sequence Read Archive), a public repository in order to ascertain the level of annotation of important sequencing steps in submissions to the database. Using SQL relational database queries to search for keywords that commonly occur in key preparatory protocol steps (fragmentation, ligation and enrichment) partitioned over studies, we found that 7.2%, 7.17% and 7.23% of all records, respectively, had at least one keyword corresponding to one of the three protocol steps. Only 4.14% of all records had keywords for all three protocol steps (5.58% of all SRA studies). Although the trend is of increasing proportion of annotation by year, the overall proportion of fully annotated records in the archive currently remains relatively low – 84.76% of all SRA studies were found to be un-annotated for these key protocol steps. We found a significant number of records were also un-dated. While the SRA provides programmatic access via an API, interoperability however is impaired by the lack of the meta-data content and the use of free-text fields. The relatively low fraction of annotation and the format it is provided in will impact how one can critically assess the quality of these biological data sets.

Real-time epidemiology and epidemic containment using social media

Jennifer Cole

There is a wealth of information on health and well-being online, provided by national government departments and agencies such as the Department of Health, Public Health England and the NHS; charities such as Diabetes UK; self-help groups initiated by patients and carers; academic institutions; the media and commercial drug companies and healthcare providers. During a Public Health Emergency of International Concern (PHEIC), such as a major influenza pandemic or serious bioterrorism attack, the availability and accuracy of such information will be vital to ensuring an effective response to the outbreak. Healthcare professionals and the public will need to know how to recognise and report symptoms; what medical help or over-the-counter drugs and medical counter-measures they need and where they are available; and how to take actions to protect themselves from catching or spreading the disease.

To date, however, there has been little research on how people search for health information online, how they choose which websites and social media feeds to trust, and how these trust validators affect the way in which they absorb and act on information. Such understanding is vital in building a robust understanding of how information that will be trusted by the public, how it will be acted on, how such channels can be used to direct appropriate behaviour, and how such information can be generated quickly during a public health emergency. Better understanding is needed on what information people want, who they want it from, why and how they choose to believe it, and what actions they will take based on the information received.

Most research into online trust validators focuses on the technological aspects of the communication channels used, overlooking the human factors that operate at the human/machine interface. Websites and social media platforms are treated as the messengers, rather than simply the medium used to deliver a message from a trusted (or untrusted) source. My research intends to focus on how and why individuals searching for health information online choose to trust information from some sources and discard information from others, including consideration of how the identity of message sender is visualised.

Furthermore, it will aim to explore the issues of information requirements, information absorption and processing, and how information received is turned into action.

Information diffusion

Andrej Žukov Gregorič

Information diffusion is a large area of research in complex networks. Given a graph, information may diffuse across its edges in a cascade-like fashion.

Cascades need not be independent of each other. Different cascades originating from different nodes may interact and thwart or magnify each other's progress. Cascades can also change the underlying graph's topology and even pave the way for new similar cascades.

Our poster focuses on four different topics: the properties of cascades, especially within online social networks; the predictability of cascades, and how to posit a robust learning problem; the modelling of cascades; and the interaction between cascades.

Cascades exhibit a number of interesting properties. The distribution of their sizes usually follows a power law distribution in online social networks. Their structure also varies, some cascades are more "viral" than others and a Wiener index can be used as a measure of their virality. We look into these and other properties of cascades.

Predicting cascade growth isn't easy. Most cascades stop after only a few infections. An interesting way of positing a learning problem is to classify whether or not a cascade will reach the median size reached by all cascades of, final, size greater than or equal to itself. Given the properties of cascades in online social networks this is analogous to asking whether the cascade will double in size. We review the existing research in this domain and present our own results.

For a long time cascades have been modelled using epidemiological disease-propagation models (e.g. SIRS), sociological models such as *Independent Cascade*, and game-theoretic models in economics. One thing to keep in mind is that the topology of the underlying graph across which cascades propagate constrains the behaviour of such models. We review information diffusion models for online social networks which have a particular structure.

It is empirically evident that cascades interact. For example, two cascades spreading similar information, originating at two different nodes in the same graph, will often collide and then compete for attention. We review the empirical evidence of such phenomenon and propose potential future work in modelling these interactions.

Category theoretical intuitions for type theory applications

Georgiana Lungu

Every day, in all domains mathematical and categorical theoretic models are being used to reason about, discover and understand properties of the various things people study. This poster discusses the alternative to the Set Theoretical Model for Natural Language, namely a Modern Type Theoretical with coercive subtyping one, and how intuitive and self contained is it. I will argue that a Categorical Theoretic model of the Type Theory that interprets Natural Language might be needed for better understanding, further development of this framework, and a more comprehensive deduction system and look at some particular details.

Parameterized directed k-Chinese postman problem and k arc-disjoint cycles problem on Euler digraphs

Bin Sheng

In the Directed k-Chinese Postman Problem (k-DCPP), we are given a connected weighted digraph G and asked to find k non-empty closed directed walks covering all arcs of G such that the total weight of the walks is minimum. Gutin, Muciaccia and Yeo (Theor. Comput. Sci. 513, 2013, 124–128) asked for the parameterized complexity of k-DCPP when k is the parameter. We prove that the k-DCPP is fixed-parameter tractable. We also consider a related problem of finding k arc-disjoint directed cycles in an Euler digraph, parameterized by k. Slivkins (ESA 2003) showed that this problem is W[1]-hard for general digraphs. Generalizing another result by Slivkins, we prove that the problem is fixed-parameter tractable for Euler digraphs. The corresponding problem on vertex-disjoint cycles in Euler digraphs remains W[1]-hard even for Euler digraphs.