Summary

Volume 17:

Workshop on Applications of Pattern Analysis

Tom Diethe, José L. Balcázar, John Shawe-Taylor, and Cristina Tîrnăucă

Preface

This is now the second edition of the Workshop on Applications of Pattern Analysis. Building on the success of the first workshop, this PASCAL supported event aims to provide a forum for applied work, and work that bridges the theory-practice gap, that sometimes is overlooked in the increasingly theoretical fields of Pattern Analysis, Computational Statistics, and Machine Learning. The workshop location for the 2011 edition was the CIEM in Castro Urdiales, Spain, and with speakers from the UK, Canada, and New Zealand, amongst others, this shows that the international appeal of such a workshop.

Pattern Analysis and Statistical Learning cover a wide range of technologies and theoretical frameworks, and significant activity in the past years has resulted in a remarkable convergence and many advances in the theory and principles underlying the field.

Bringing these technologies to real world demanding applications is however often treated as a separate problem, one that does not directly affect the field as a whole. It is instead important to consider the field of Pattern Analysis as fully including all issues involved with the applications of this technology, and hence all issues that arise when deploying, scaling, implementing and using the technology.

The workshop called for contributions in the form of Demos, Case Studies, Working Systems, Real World Applications and Usage Scenarios. Challenges may stem from the violation of common theoretical assumptions, from the specific types of patterns and noise arising in certain scenarios, or from the problem of scaling up the implementation of state of the art algorithms to real world sizes, or from the creation of integrated software systems that contain multiple pattern-analysis components.

We were also interested in new application areas, where Pattern Analysis has been deployed with success, and in issues involving the visualisation and delivery and exploitation of the patterns discovered by PA technologies. Systems working in noisy and unstructured environments and situations are particularly interesting.

The goal was to discuss and reward work aimed at making theory useful and relevant, without requesting the researchers to propose new theoretical methods, but rather requesting to show how they solved the many challenges related to applying these methods to real world scenarios, or how they benefited other fields of research. Getting ideas to work in real scenarios is what this is about.

We believe that the papers are of a high standard and meet the goals of the workshop. In addition to the submissions, gathered here, the program of the workshop featured two invited talks and a whole afternoon session devoted to projects from the Harvest Programme of the Pascal-2 Network of Excellence; in order to give a more faithful description here, we include the abstracts of both invited talks and a brief description of the Harvest session.

Invited Talks

1. Utilizing unlabeled and weakly labeled samples in classification learning tasks
   Shai Ben-David, University of Waterloo

   Abstract: *In many classification learning tasks, labeled data may be expensive or scarce. At the same time, unlabeled or "weakly labeled" samples, may be available in abundance. We consider three algorithmic paradigms that utilize unlabeled or "weakly labeled" samples to help classification tasks. On top of proposing some meta-algorithms for utilizing such samples, we analyse the sample complexity of these paradigms. We show that in some semi-supervised learning task, as well as in some domain adaptation and query learning tasks, unlabeled samples can be applied to provably achieve saving in the sizes of required labeled samples.*

2. Medical Text Mining
   Tom Diethe, British Medical Association

   Abstract: *The British Medical Journal Group (BMJ Group) has a wide and varied content set, including a suite of medical journals, online learning materials, best practice guidelines, clinical evidence summaries, a doc-2-doc online forum, and a portfolio system for doctors. There is an emerging need to aggregate accross these content types, providing a unified tagging and linking system, so that related content can easily be retrieved across the group. The main use-cases include an improved search and browse capability, and the (semi-)automatic construction of "specialty portals", which may be medical in nature (e.g. diabetes) or non-medical (e.g. NHS reform). This provides a challenge to standard Pattern Analysis algorithms, due in part to the highly technical nature of the documents. Prior work has mainly been focussed on the use of tools that automatically index against a medical ontology (such as MetaMap and UMLS), but this approach has drawbacks in terms of computational resources, lack of user control, and limitations to medical-only concepts. A hybrid approach based on statistiscal and semantic methods appears to have some merit, and may be the way forward. The presentation will focus on preliminary work taking the two approaches, and talk about some specific technical issues that have arisen along the way. This is based on joint work with Jonathon Peterson, Chris Wroe, and Rob Challen.*

Harvest Programme Session

The Harvest Programme is one of the research funding lines of the Pascal-2 Network of Excellence of the 7th Framework Programme of the European Union. Harvest Projects have some piece of software as their main objective. Teams are expected to be mixed, with some members coming from academia/public research and others coming from industry or from a field outside the direct scope of PASCAL 2. These projects are expected to expose their results at one or more workshops or conferences. The Harvest session of WAPA 2011 reports on two such projects: one oriented to extend an open-source natural language processing system, Freeling, with structured prediction capabilities; the other aims at reimplementing and improving a recent association miner, yacaree, designed at a Pascal-2 site, as a node

of the open-source data mining tool KNIME, widely used in commercial data mining and supported by a Swiss company. The session included a presentation of the KNIME tool by a member of the KNIME team.

October 2011

The Editorial Team:

Tom Diethe
University College London
t.diethe@cs.ucl.ac.uk

José L. Balcázar
Universidad de Cantabria
joseluis.balcazar@unican.es

John Shawe-Taylor
University College London
J.Shawe-Taylor@cs.ucl.ac.uk

Cristina Tîrnăucă
Universidad de Cantabria
cristina.tirnauca@unican.es