Mach Learn (2010) 81: 1–4
DOI 10.1007/s10994-010-5200-0

EDITORIAL

Special issue for ECML PKDD 2010: Guest editors' introduction

José L Balcázar · Francesco Bonchi · Aristides Gionis · Michèle Sebag

Received: 30 April 2010 / Accepted: 20 June 2010 / Published online: 30 July 2010

This special issue of the Machine Learning Journal presents selected papers from the 2010 edition of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery from Databases (ECML PKDD). Since the early 2000s ECML PKDD has become the premier scientific event worldwide gathering researchers in both fields of Machine Learning and Data Mining, allowing them to benefit from, and contribute to, progress in the sibling field.

In 2010 ECML PKDD attracted 658 full-paper submissions, which underwent a rigorous reviewing process. Each paper was assigned to three reviewers and one Area Chair. The selection, made on the basis of novelty, creativity, scientific quality and overall impact of the work to the community, led to 120 papers to be presented in the ECML PKDD conference.

Out of these 120 papers, seven papers have been selected to form this special issue of the Machine Learning Journal – for their exceptional quality and their potential to spur new research in the field.

J.L Balcázar
Dept. Matemáticas, Estadística y Computación, Universidad de Cantabria Santander, Santander, Spain
e-mail: joseluis.balcazar@unican.es

F. Bonchi (✉) · A. Gionis
Yahoo! Research Barcelona, Avinguda Diagonal 177, 08018 Barcelona, Spain
e-mail: bonchi@yahoo-inc.corp

A. Gionis
e-mail: gionis@yahoo-inc.corp

M. Sebag
TAO, CNRS – INRIA – LRI, Université Paris-Sud, 91405 Orsay, France
e-mail: sebag@lri.fr

Language and vision

Being able to deal with texts and natural language at large appears as a key challenge, for most of the existing knowledge is expressed through language, and using language is one of the most characteristic human activities. How to understand a text, how to extract information from corpora, and knowledge from information, has been aimed at for over three decades in Machine Learning and Computational Linguistics. On the one hand, this long term goal has been decomposed into more actionable tasks, ranging from Named Entity Recognition to e.g., Textual Entailment and Question/Answering. On the other hand, ever more comprehensive generative and/or discriminant models have been proposed to account for the complex texture of language. The optimal solution would of course enforce some trade-off between induction and deduction (what can be learned from the available evidence; how to use it when learned), according to the general "Learning to Reason" principles. The first paper in this volume, *A Segmented Topic Model based on the Two-parameter Poisson-Dirichlet Process* by Lan Du, Wray Buntine and Huidong Jin, specifically addresses this issue. The authors model the interleaved structure of topics and words within a document through a Two-parameter Poisson Dirichlet process, thus achieving a novel and promising tradeoff between frugal estimation and powerful inference via a hierarchical model, sharing the distribution estimates among the different segments of the document.

Another key vehicle of information is by means of images. The automatic processing of images naturally raises critical issues, too; one specific difficulty in comparison to text understanding comes from the fact that one pixel on its own virtually conveys no information, contrasting with one word. How to build relevant higher-level features for Image Understanding is at the core of the Pattern Recognition field. Image annotation could actually be viewed as one way of building such features, as mapping an image into a (possibly weighted) set of words will leverage text understanding into image understanding. An even more ambitious goal is to tame the joint structure of images and annotations, possibly overcoming the multi-lingual issue. Scalability is a "must-have" feature here, for two reasons. On the one hand, one lesson seemingly learned from many text and image-related challenges in the recent years is that simple processing of comprehensive data collection outperforms complex processing on restricted data collection (active learning excluded). On the other hand, the World-Wide Web makes it possible to collect virtually infinite data for free. The second paper in this volume, *Large Scale Image Annotation: Learning to Rank with Joint Word-Image Embeddings* by Jason Weston, Samy Bengio and Nicolas Usunier, investigates how to map images and annotations onto a single low-dimensional space, such that related (image,annotations) are mapped near each other. Scalability and robustness are achieved by applying an online, stochastic gradient-based approach to a top-k rank-based criterion.

Relations, spectral analysis and random walks

Uncovering the structure of the domain from relations among examples and entities is at the root of several Machine Learning approaches with rich theoretical foundations, including Support Vector Machines, Spectral Analysis and Relational Learning. Spectral Analysis usually starts from a matrix describing the flow of information among examples, e.g. the graph Laplacian. The challenge is to recover the underlying structure of the problem domain from the eigenvectors of the matrix, e.g. harnessing max-cut algorithms to achieve optimal clustering. The third paper in this volume, *On the Eigenvectors of p-Laplacian* by Dijun Luo, Chris Ding, Heng Heng, and Feiping Nie, advances the state of the art in Spectral Analysis

both theoretically and algorithmically. Theoretically, the authors provide a full eigenvector analysis of p-Laplacian, thus obtaining a natural global embedding for multi-class clustering problems (as opposed to, recursively using a two-class clustering). Algorithmically, they provide a gradient descent optimization approach which is guaranteed to terminate in a finite number of time steps, showing empirically that this finite number of steps is reasonable in practice.

The fourth paper in this volume, *Relational Retrieval Using a Combination of Path-Constrained Random Walks* by Ni Lao and William Cohen, is also concerned with domain and example modelling in terms of graphs. This work nicely connects random walks on a graph with a rich relational setting, focussing on relational retrieval as opposed to statistical relational learning. More specifically, starting from a huge graph the goal is to find in an efficient and scalable manner the nodes most relevant to a given query, through following the appropriate edges, and weighting and aggregating the constructed paths in an optimal way. The approach indeed revisits the key tasks in relational learning and inductive logic programming: parameter estimation, structure learning, and predicate invention. As argued by the authors, the proposed approach provides a scalable inference method (as opposed to e.g. variational methods used in Markov logic networks); compared to relational kernels, the difference might be on the learning granularity (contrasting e.g. multiple kernel learning with predicate invention). Besides information retrieval in a large biological gene network, the paper considers three tasks very relevant to most scientists: expert finding (finding a reviewer), reference recommendation (which paper to cite) and venue recommendation (where to submit a paper). The guest editors can witness indeed that the expert finding functionality could dramatically simplify the task of conference chairs…

Adversarial learning and exploration

While mainstream Machine Learning routinely assumes that the training set and the test set are independent identically distributed samples, and rely on the same distribution, quite a few interesting settings that violate these assumptions have been investigated in the last decade, such as Active Learning or Multi-task Learning, to name a few. The fifth paper in this volume, *Mining Adversarial Patterns via Regularized Loss Minimization* by Wei Liu and Sanjay Chawla, considers the particular case of an *adversarial* test set, motivated by an adaptive spam filter application. In such a context indeed, the spammer might come to know the previous sample and thus the filter hypothesized by the Machine Learning algorithm; his goal thus becomes to produce spams that would pass for non-spam mails w.r.t. this hypothesis. The contribution of the paper is to formalize adversarial learning as a zero-sum game, specifically a Stackelberg game, where the filter and the spammer alternatively learn a hypothesis from the current sample, and perturb the sample (generating deceptive examples) in order to defeat the current hypothesis. Both players are assumed to be "rational" in the sense that the spammer will generate the optimally deceptive sample conditionally to the current hypothesis while the filter will learn the optimal (linear) hypothesis conditionally to the current sample. This work is thus relevant to robust learning, reaching the equilibrium strategy through a convex optimization procedure involved with a regularization term on the filter side, and a "perturbation" cost on the spammer side.

Reinforcement Learning is another Machine Learning setting where gathering examples from the environment is part of the learning task. In the sixth paper of this volume entitled *Dimension Reduction and Its Application to Model-based Exploration in Continuous Spaces*, Ali Nouri and Michael Littman focus on model-based Reinforcement Learning,

where the environment is described as a continuous state space. A salient feature of the proposed approach is to identify the system dynamics (the transition function) using a multivariate kernel regression algorithm combined with dimensionality reduction. The sample complexity of the approach is significantly reduced through dimensionality reduction on the one hand, and because the exploration is guided from the quality of the current estimate (akin to active learning) on the other hand. While the identification of the "most unknown regions" in the initial environment representation admittedly is the most computationally demanding part of the approach, a warm restart strategy enforces computationally efficient results. Interestingly, the integration of Active Learning within Reinforcement Learning also is at the core of Autonomous Robotics and Embodied Statistical Learning, e.g. defining self-driven rewards ("curiosity").

When everything else fails, look for priors: a geometric approach

The last paper of this volume, *A Geometric View of Conjugate Priors* by Arvind Agarwal and Hal Daumé III, revisits the magic of Bayesian approaches, and how to get going through providing prior knowledge in a hybrid generative/discriminative framework. Provokingly, the authors suggest that priors should not be a nickname for mathematical or computational convenience. Within an exponential family framework, they discuss the geometry induced by different prior distributions in terms of Bregman divergence. The argument goes as "it is reasonable to measure Bregman divergence between real data points and the parameters to estimate, in the same way as between the pseudo data introduced by the prior and the parameters". As summarized by one reviewer (fairly reflecting the opinion of the other two reviewers): *This is a jolly good paper!*

We wish to warmly thank all authors of the submitted papers for their contribution to the quality of ECML PKDD. We are also very grateful to all Area Chairs, to all reviewers in the Program Committee and additional reviewers, for the very hard and intense selection work distilled in this special issue.

Enjoy reading.