2nd Asian Conference on Machine Learning (ACML2010)
Tokyo, Japan, November 8–10, 2010

http://sugiyama-www.cs.titech.ac.jp/ACML2010/

Preface

The 2nd Asian Conference on Machine Learning (ACML2010) was held on November 8–10, 2010, at Tokyo Institute of Technology in Tokyo, Japan. The conference aims at providing a leading international forum for researchers in machine learning and related fields to share their new ideas and achievements. The conference called for research papers reporting original investigation results and proposals focusing on frontier research in all aspects of machine learning.

Although the conference is named "Asian Conference", we solicited submissions also from other than the Asia-Pacific regions. We have received 74 submissions from all over the world. Each paper was carefully reviewed by 4 program committee members and a senior program committee member, and we finally decided to accept 22 papers for oral presentation, with acceptance rate 29.7%. Although there was no explicit control for geographical balancing, the accepted papers were truly coming from all over the world. The paper review process involved 19 senior program committee members and 127 program committee members, all of whom volunteered their valuable time to provide high-quality reviews despite the tight reviewing schedule.

ACML2010 was a single-track conference, and all the 22 accepted papers were presented sequentially on November 9th and 10th, 2010. The conference also featured 3 tutorials on November 8th, 3 invited talks on November 9th and 10th, and a poster session consisting of 19 posters on November 10th, 2010.

We would like to acknowledge the Air Force Office of Scientific Research, Asian Office of Aerospace Research and Development (AFOSR/AOARD), and the Office of Naval Research Global (ONRG) for their financial support, which was mainly used for supporting students' participation. We acknowledge the general chair Takashi Washio and ACML steering committee members for their advice on conference organization, Hirotaka Hachiya for managing the conference web pages, and Hiroko Iida of Keio Travel Agency, Co., LTD. for making arrangement for travel and accommodation. We also thank Tokyo Institute of Technology, Global COE: Computationism as a Foundation for the Sciences, JST PRESTO: Synthesis of Knowledge for Information Oriented Society, IEICE Technical Group on Information-Based Induction Sciences and Machine Learning (IBISML)

for their support. Our special thanks also go to Microsoft Research for allowing us to use the Conference Management Toolkit for preparing the conference. Finally, we are grateful to all the conference participants and those who submitted their papers to the conference.

Program Committee Co-chairs
Masashi Sugiyama (Tokyo Institute of Technology, Japan)
Qiang Yang (Hong Kong University of Science and Technology, China)

Conference Program

Nov. 8 (Day 1)

08:30–10:00 Tutorial 1 (Chaired by Masashi Sugiyama)

- *Web People Search: Person Name Disambiguation and Other Problems*, Minoru Yoshida and Hiroshi Nakagawa

10:00–10:20 Coffee Break

10:20–11:20 Tutorial 1 (Chaired by Masashi Sugiyama)

- *Web People Search: Person Name Disambiguation and Other Problems*, Minoru Yoshida and Hiroshi Nakagawa

11:20–11:30 Break

11:30–12:30 Tutorial 2 (Chaired by Thomas G. Dietterich)

- *Honest Evaluation of Classification Models*, Jose A. Lozano, Guzman Santafe, and Inaki Inza

12:30–14:00 Lunch Break (on Your Own)

14:00–15:30 Tutorial 2 (Chaired by Thomas G. Dietterich)

- *Honest Evaluation of Classification Models*, Jose A. Lozano, Guzman Santafe, and Inaki Inza

15:30–15:50 Coffee Break

15:50–16:50 Tutorial 3 (Chaired by Koji Tsuda)

- *Support Vector Machines and Kernel Methods: Status and Challenges*, Chih-Jen Lin

16:50–17:00 Break

17:00–18:00 Tutorial 3 (Chaired by Koji Tsuda)

- *Support Vector Machines and Kernel Methods: Status and Challenges*, Chih-Jen Lin

Nov. 9 (Day 2)

08:45-09:00	Opening
09:00–10:00	Invited Talk 1 (Chaired by Qiang Yang) • *Optimal Online Prediction in Adversarial Environments*, Peter L. Bartlett
10:00–10:20	Coffee Break
10:20–11:40	Session 1: Statistical Learning (Chaired by Zhi-Hua Zhou) • *Pairwise Measures of Causal Direction in Linear Non-Gaussian Acyclic Models*, Aapo Hyvarinen • *Learning Polyhedral Classifiers Using Logistic Function*, Naresh Manwani and P. S. Sastry • *Ellipsoidal Support Vector Machines*, Michinari Momma, Kohei Hatano, and Hiroki Nakayama • *Minimum Conditional Entropy Clustering: A Discriminative Framework for Clustering*, Bo Dai and Baogang Hu
11:40–13:30	Lunch Break (on Your Own)

(continue to the next page)

13:30–14:50	Session 2: Bayesian Learning (Chaired by Seungjin Choi) • *Efficient Collapsed Gibbs Sampling for Latent Dirichlet Allocation*, Han Xiao and Thomas Stibor • *Variational Relevance Vector Machine for Tabular Data*, Dmitry Kropotov, Dmitry Vetrov, Lior Wolf, and Tal Hassner • *Hierarchical Gaussian Process Regression*, Sunho Park and Seungjin Choi • *Content-based Image Retrieval with Multinomial Relevance Feedback*, Dorota Glowacka and John Shawe-Taylor
14:50–15:00	Break
15:00–16:00	Invited Talk 2 (Chaired by Takashi Washio) • *Learning without Search*, Geoff Webb
16:00–16:20	Coffee Break
16:20–18:00	Session 3: Discretization, Logic, Graphs, and Rules (Chaired by Yuji Matsumoto) • *The Coding Divergence for Measuring the Complexity of Separating Two Sets*, Mahito Sugiyama and Akihiro Yamamoto • *Single versus Multiple Sorting in All Pairs Similarity Search*, Yasuo Tabei, Takeaki Uno, Masashi Sugiyama, and Koji Tsuda • *An EM Algorithm on BDDs with Order Encoding for Logic-based Probabilistic Models*, Masakazu Ishihata, Yoshitaka Kameya, Taisuke Sato, and Shin-ichi Minato • *Exploiting the High Predictive Power of Multi-class Subgroups*, Tarek Abudawood and Peter Flach • *Generative Models of Information Diffusion with Asynchronous Time-delay*, Kazumi Saito, Masahiro Kimura, Kouzou Ohara, and Hiroshi Motoda
18:30–20:30	Banquet

Nov. 10 (Day 3)

09:00–10:00	Invited Talk 3 (Chaired by Masashi Sugiyama) • *Kernel Method for Bayesian Inference*, Kenji Fukumizu
10:00–10:20	Coffee Break
10:20–11:40	Session 4: Stream and Large-scale Data (Chaired by Chih-Jen Lin) • *Accurate Ensembles for Data Streams: Combining Restricted Hoeffding Trees using Stacking*, Albert Bifet, Eibe Frank, Geoffrey Holmes, and Bernhard Pfahringer • *Mining Recurring Concept Drifts with Limited Labeled Streaming Data*, Peipei Li, Xindong Wu, and Xuegang Hu • *Hierarchical Convex NMF for Clustering Massive Data*, Kristian Kersting, Mirwaes Wahabzada, Christian Thurau, and Christian Bauckhage • *Multi-task Learning for Recommender System*, Xia Ning and George Karypis
11:40–13:30	Lunch Break

(continue to the next page)

13:30–14:50	Session 5: Reinforcement Learning (Chaired by Remi Munos)

- *Adaptive Step-size Policy Gradients with Average Reward Metric*, Takamitsu Matsubara, Tetsuro Morimura, and Jun Morimoto
- *Finite-sample Analysis of Bellman Residual Minimization*, Odalric-Ambrym Maillard, Remi Munos, Alessandro Lazaric, and Mohammad Ghavamzadeh
- *A Study of Approximate Inference in Probabilistic Relational Models*, Fabian Kaelin and Doina Precup
- *Conceptual Imitation Learning: An Application to Human-robot Interaction*, Hossein Hajimirsadeghi, Majid Nili Ahmadabadi, Mostafa Ajallooeian, Babak Araabi, and Hadi Moradi

14:50–15:00	Break
15:00–15:40	Poster Spotlight Session (Chaired by Masashi Sugiyama)
15:40–16:00	Coffee Break
16:00–18:00	Poster Session
18:00–18:10	Closing

Invited Talks

1. Optimal Online Prediction in Adversarial Environments
   Peter L. Bartlett (UC Berkeley)

   In many prediction problems, including those that arise in computer security and computational finance, the process generating the data is best modeled as an adversary with whom the predictor competes. The predictor's aim is to minimize the regret, or the difference between the predictor's performance and the best performance among some comparison class, whereas the adversary aims to maximize the predictor's regret. Even decision problems that are not inherently adversarial can be usefully modeled in this way, since the assumptions are sufficiently weak that effective prediction strategies for adversarial settings are very widely applicable.

   The first part of this talk presents an example of online decision problems of this kind: a resource allocation problem from computational finance. We describe an efficient strategy with near-optimal performance.

   The second part of the talk presents results on the regret of optimal strategies. These results are closely related to finite sample analyses of prediction strategies for probabilistic settings, where the data are chosen iid from an unknown probability distribution. In particular, we show that the optimal online regret is closely related to the behavior of empirical minimization in a probabilistic setting, but with a non-iid stochastic process generating the data. This allows the application of techniques from the analysis of the performance of empirical minimization in an iid setting, which relates the optimal regret to a measure of complexity of the comparison class that is similar to the Rademacher averages that have been studied in the iid setting.

   Biography: Peter Bartlett is a professor in the Computer Science Division and the Department of Statistics at the University of California at Berkeley. He is the co-author, with Martin Anthony, of the book Learning in Neural Networks: Theoretical Foundations, has edited three other books, and has co-authored many papers in the areas of ma-

chine learning and statistical learning theory. He has served as an associate editor of the journals Machine Learning, Mathematics of Control Signals and Systems, the Journal of Machine Learning Research, the Journal of Artificial Intelligence Research, and the IEEE Transactions on Information Theory, as a member of the editorial boards of Machine Learning, the Journal of Artificial Intelligence Research, and Foundations and Trends in Machine Learning, and as a member of the steering committees of the Conference on Computational Learning Theory and the Algorithmic Learning Theory Workshop. He has consulted to a number of organizations, including General Electric, Telstra, Polaris Wireless and SAC Capital Advisors. In 2001, he was awarded the Malcolm McIntosh Prize for Physical Scientist of the Year in Australia, for his work in statistical learning theory. He was a Miller Institute Visiting Research Professor in Statistics and Computer Science at U.C. Berkeley, a fellow, senior fellow and professor in the Research School of Information Sciences and Engineering at the Australian National University's Institute for Advanced Studies, and an honorary professor in the School of Information Technology and Electrical Engineering at the University of Queensland. His research interests include machine learning, statistical learning theory, and adaptive control.

2. Learning without Search
   Geoff Webb (Monash University)

   Machine learning is classically conceived as search through a hypothesis space for a hypothesis that best fits the training data. In contrast, naive Bayes performs no search, extrapolating an estimate of a high-order conditional probability by composition from lower-order conditional probabilities. In this talk I show how this searchless approach can be generalised, creating a family of learners that provide a principled method for controlling the bias/variance trade-off. At one extreme very low variance can be achieved as appropriate for small data. Bias can be decreased with larger data in a manner that ensure Bayes optimal asymptotic error. These algorithms have the desirable properties of

   - training time that is linear with respect to training set size,

- supporting parallel and anytime classification,
- allowing incremental learning,
- providing direct prediction of class probabilities,
- supporting direct handling of missing values, and
- robust handling of noise.

Despite being generative, they deliver classification accuracy competitive with state-of-the-art discriminative techniques.

Biography: Geoff Webb holds a research chair in the Faculty of Information Technology at Monash University, where he heads the Centre for Research in Intelligent Systems. Prior to Monash he held appointments at Griffith University and then Deakin University, where he received a personal chair. His primary research areas are machine learning, data mining, and user modelling. He is known for the development of numerous methods, algorithms and techniques for machine learning, data mining and user modelling. His commercial data mining software, Magnum Opus, incorporates many techniques from his association discovery research. Many of his learning algorithms are included in the widely-used Weka machine learning workbench. He is editor-in-chief of the highest impact data mining journal, Data Mining and Knowledge Discovery, co-editor of the Encyclopedia of Machine Learning (to be published by Springer), a member of the advisory board of Statistical Analysis and Data Mining and a member of the editorial boards of Machine Learning and ACM Transactions on Knowledge Discovery in Data.

3. Kernel Method for Bayesian Inference
   Kenji Fukumizu (The Institute of Statistical Mathematics)

   Since the proposal of support vector machine, various kernel methods have been extensively developed as nonlinear extensions or "kernelization" of classical linear methods. More recently, however, it has become clear that a potentially more reaching use of kernels is a linear way of dealing higher order statistics by embedding distributions as the form of means in reproducing kernel Hilbert spaces (RKHS) and by considering linear operators among them.

   This talk will present how general Bayesian inference can be realized based on this recent recognition of the kernel method. First, I will explain the kernel method for expressing conditional probabilities by the kernel covariance operators of the distributions. Second, it will be shown that the general Bayes' rule, which is the center of Bayesian inference, is realized by operations on the kernel expression of the conditional probability and the prior represented as the mean in RKHS. The kernel mean of the posterior is obtained by Gram matrix computations to realize the procedure of Bayes' rule: constructing the joint probability and its normalization. The rate of convergence of the empirical kernel estimate to the true posterior is also derived.

As an application, I will discuss kernel nonparametric HMM, in which the conditional probabilities to define the HMM model are neither given in a specific form nor estimated with a parametric model, but given in the form of finite samples. By sequential application of the kernel Bayes' rule, it will be shown with some experiments that the hidden states can be sequentially estimated nonparametrically.

Biography: Kenji Fukumizu is a professor in the Department of Statistical Modeling at The Institute of Statistical Mathematics, where he serves as director of the Research Innovation Center. Prior to the current institute, he worked as a researcher in the Research and Development Center, Ricoh Co., Ltd. and the Institute of Physical and Chemical Research (RIKEN). He was a visiting scholar at the Department of Statistics, UC Berkeley, and a Humboldt fellow at Max Planck Institute for Biological Cybernetics. He serves as an associate editor of the journals, Annals of the Institute of Statistical Mathematics, Neural Networks, and Foundations and Trends in Machine Learning. His research interests include machine learning and mathematical statistics. He has co-authored a book on singular statistical models, and has authored a book on kernel methods (to be published in 2010).

Tutorials

1. Web People Search: Person Name Disambiguation and Other Problems
   Minoru Yoshida(University of Tokyo) and Hiroshi Nakagawa (University of Tokyo)

   This tutorial will present the current state of research on Web people searches. It will be mainly about person name disambiguation problems, and the attribute extraction methods that can be used to support person name disambiguation. We shall give a survey of the algorithms and tools available, and discuss the possibility of applying machine learning to this problem. A survey of WePS workshops dedicated to this task will also be presented.

2. Honest Evaluation of Classification Models
   Jose A. Lozano (University of the Basque Country), Guzman Santafe (University of the Basque Country), and Inãki Inza (University of the Basque Country)

   The objective of the tutorial is to give an overview on validation methods of supervised classification algorithms. The tutorial starts by presenting the most common performance measures used to evaluate supervised learning algorithm. After that the methods used to estimate the previous measures will be described in detail. We will also expose the statistical tests that can be used to compare several supervised classification algorithms. The tutorial concludes by giving recommendations to perform honest classifier evaluation according to specific characteristics of the problem or the data set at hand as well as general best practices in classifier evaluation.

3. Support Vector Machines and Kernel Methods: Status and Challenges
   Chih-Jen Lin (National Taiwan University, Taiwan)

   Support vector machines (SVM) and kernel methods are now important machine learning techniques. In this tutorial, we first introduce some basic concepts such as maximal margin, kernel mappings, and primal dual relationships. We then discuss the training by solving optimization problems and the selection of parameters. Finally, we briefly mention some new research issues.

Accepted Papers

1. Pairwise Measures of Causal Direction in Linear Non-Gaussian Acyclic Models
   Aapo Hyvarinen (University of Helsinki)

   We present new measures of the causal direction between two non-gaussian random variables. They are based on the likelihood ratio under the linear non-gaussian acyclic model (LiNGAM). We also develop simple first-order approximations and analyze them based on related cumulant-based measures. The cumulant-based measures can be shown to give the right causal directions, and they are statistically consistent even in the presence of measurement noise. We further show how to apply these measures to estimate LiNGAM for more than two variables, and even in the case of more variables than observations. The proposed framework is statistically at least as good as existing ones in the cases of few data points or noisy data, and it is computationally and conceptually very simple.

2. Learning Polyhedral Classifiers Using Logistic Function
   Naresh Manwani (Indian Institute of Science) and P. S. Sastry (Indian Institute of Science)

   In this paper we propose a new algorithm for learning polyhedral classifiers. In contrast to existing methods for learning polyhedral classifier which solve a constrained optimization problem, our method solves an unconstrained optimization problem. Our method is based on a logistic function based model for the posterior probability function. We propose an alternating optimization algorithm, namely, SPLA1 (Single Polyhedral Learning Algorithm1) which maximizes the log-likelihood of the training data to learn the parameters. We also extend our method to make it independent of any user specified parameter (e.g., number of hyperplanes required to form a polyhedral set) in SPLA2. We show the effectiveness of our approach with experiments on various synthetic and real world datasets and compare our approach with a standard decision tree method (OC1) and a constrained optimization based method for learning polyhedral sets.

3. Ellipsoidal Support Vector Machines
   Michinari Momma (NEC), Kohei Hatano (Kyushu University), and Hiroki Nakayama (NEC)

   This paper proposes the ellipsoidal SVM (e-SVM) that uses an ellipsoid center, in the version space, to approximate the Bayes point. Since SVM approximates it by a sphere center, e-SVM provides an extension to SVM for better approximation of the Bayes point. Although the idea has been mentioned before (Rujan, 1997), no work has been done for formulating and kernelizing the method. Starting from the maximum volume ellipsoid problem, we successfully formulate and kernelize it by employing relaxations. The resulting e-SVM optimization framework has much similarity to SVM; it is naturally extendable to other loss functions and other problems. A variant of the sequential minimal optimization is provided for efficient batch implementation. Moreover, we provide an online version of linear, or primal, e-SVM to be applicable for large-scale datasets.

4. Minimum Conditional Entropy Clustering: A Discriminative Framework for Clustering
   Bo Dai (NLPR/LIAMA) and Baogang Hu (NLPR/LIAMA)

   In this paper, we introduce an assumption which makes it possible to extend the learning ability of discriminative model to unsupervised setting. We propose an information-theoretic framework as an implementation of the low-density separation assumption. The proposed framework provides a unified perspective of Maximum Margin Clustering (MMC), Discriminative k-means, Spectral Clustering and Unsupervised Renyi's Entropy Analysis and also leads to a novel and efficient algorithm, Accelerated Maximum Relative Margin Clustering (ARMC), which maximizes the margin while considering the spread of projections and affine invariance. Experimental results show that the proposed discriminative unsupervised learning method is more efficient in utilizing data and achieves the state-of-the-art or even better performance compared with mainstream clustering methods.

5. Efficient Collapsed Gibbs Sampling for Latent Dirichlet Allocation
   Han Xiao (Technical University of Munich) and Thomas Stibor (Technical University of Munich)

Collapsed Gibbs sampling is a frequently applied method to approximate intractable integrals in probabilistic generative models such as latent Dirichlet allocation. This sampling method has however the crucial drawback of high computational complexity, which makes it limited applicable on large data sets. We propose a novel dynamic sampling strategy to significantly improve the efficiency of collapsed Gibbs sampling. The strategy is explored in terms of efficiency, convergence and perplexity. Besides, we present a straight-forward parallelization to further improve the efficiency. Finally, we underpin our proposed improvements with a comparative study on different scale data sets.

6. Variational Relevance Vector Machine for Tabular Data
   Dmitry Kropotov (Dorodnicyn Computing Centre), Dmitry Vetrov (Lomonosov Moscow State University), Lior Wolf (Tel Aviv University), and Tal Hassner (The Open University of Israel)

   We adopt the Relevance Vector Machine (RVM) framework to handle cases of table-structured data such as image blocks and image descriptors. This is achieved by coupling the regularization coefficients of rows and columns of features. We present two variants of this new gridRVM framework, based on the way in which the regularization coefficients of the rows and columns are combined. Appropriate variational optimization algorithms are derived for inference within this framework. The consequent reduction in the number of parameters from the product of the table's dimensions to the sum of its dimensions allows for better performance in the face of small training sets, resulting in improved resistance to overfitting, as well as providing better interpretation of results. These properties are demonstrated on synthetic data-sets as well as on a modern and challenging visual identification benchmark.

7. Hierarchical Gaussian Process Regression
   Sunho Park (POSTECH) and Seungjin Choi (POSTECH)

   We address an approximation method for Gaussian process (GP) regression, where we approximate covariance by a block matrix such that diagonal blocks are calculated exactly while off-diagonal blocks are approximated. Partitioning input data points, we present a two-layer hierarchical model for GP regression, where prototypes of clusters in the upper layer are involved for coarse modeling by a GP and data points in each cluster in the lower layer are involved for fine modeling by an individual GP whose prior mean is given by the corresponding prototype and covariance is parameterized by data points in the partition. In this hierarchical model, integrating out latent variables in the upper layer leads to a block covariance matrix, where diagonal blocks contain similarities between data points in the same partition and off-diagonal blocks consist of approximate similarities calculated using prototypes. This particular structure of the covariance matrix divides the full GP into a pieces of manageable sub-problems whose complexity scales with the number of data points in a partition. In addition, our hierarchical GP regression (HGPR) is also useful for cases where partitions of data reveal different characteristics. Experiments on several benchmark datasets confirm the useful behavior of our method.

8. Content-based Image Retrieval with Multinomial Relevance Feedback
   Dorota Glowacka (University College London) and John Shawe-Taylor (University College London)

   The paper considers an interactive search paradigm in which at each round a user is presented with a set of k images and is required to select one that is closest to her target. Performance is measured by the number of rounds needed to identify a specific target image or to find an image among the t nearest neighbours to the target in the database. Building on earlier work we assume a multinomial user model with the probabilities of response proportional to a function of the distance to the target. The conjugate prior Dirichlet distribution is used to model the problem motivating an algorithm that trades exploration and exploitation in presenting the images in each round. Experimental results verify the fit of the model with the problem as well as show that the new approach compares favourably with previous work.

9. The Coding Divergence for Measuring the Complexity of Separating Two Sets
   Mahito Sugiyama (Kyoto University) and Akihiro Yamamoto (Kyoto

University)

In this paper we integrate two essential processes, discretization of continuous data and learning of a model that explains them, towards fully computational machine learning from continuous data. Discretization is fundamental for machine learning and data mining, since every continuous datum; e.g., a real-valued datum obtained by observation in the real world, must be discretized and converted from analog (continuous) to digital (discrete) form to store in databases. However, most machine learning methods do not pay attention to the situation; i.e., they use digital data in actual applications on a computer whereas assume analog data (usually vectors of real numbers) theoretically. To bridge the gap, we propose a novel measure of the difference between two sets of data, called the coding divergence, and unify two processes discretization and learning computationally. Discretization of continuous data is realized by a topological mapping (in the sense of mathematics) from the d-dimensional Euclidean space $\mathbb{R}^d$ into the Cantor space Σ^ω, and the simplest model is learned in the Cantor space, which corresponds to the minimum open set separating the given two sets of data. Furthermore, we construct a classifier using the divergence, and experimentally demonstrate robust performance of it. Our contribution is not only introducing a new measure from the computational point of view, but also triggering more interaction between experimental science and machine learning.

10. Single versus Multiple Sorting in All Pairs Similarity Search
Yasuo Tabei (JST Minato ERATO Project), Takeaki Uno (National Institute of Informatics of Japan), Masashi Sugiyama (Tokyo Institute of Technology), and Koji Tsuda (National Institute of Advanced Industrial Science and Technology)

To save memory and improve speed, vectorial data such as images and signals are often represented as strings of discrete symbols (i.e., sketches). Chariker (2002) proposed a fast approximate method for finding neighbor pairs of strings by sorting and scanning with a small window. This method, which we shall call " single sorting ", is applied to locality sensitive codes and prevalently used in speed-demanding web-related applications. To improve on single sorting, we propose a novel method that employs blockwise masked sorting. Our method can dramatically reduce the number of candidate pairs which have to be verified by distance calculation in exchange with an increased amount of sorting operations. So it is especially attractive for high dimensional dense data, where distance calculation is expensive. Empirical results show the efficiency of our method in comparison to single sorting and recent fast nearest neighbor methods.

11. An EM Algorithm on BDDs with Order Encoding for Logic-based Probabilistic Models
    Masakazu Ishihata (Tokyo Institute of Technology), Yoshitaka Kameya (Tokyo Institute of Technology), Taisuke Sato (Tokyo Institute of Technology), and Shin-ichi Minato (Hokkaido University)

    Logic-based probabilistic models (LBPMs) enable us to handle various problems in the real world thanks to the expressive power of logic. However, most of LBPMs have restrictions to realize efficient probability computation and learning. We propose an EM algorithm working on BDDs with order encoding for LBPMs. A notable advantage of our algorithm over existing approaches is that it copes with multi-valued random variables without restrictions. The complexity of our algorithm is proportional to the size of the BDD. In the case of hidden Markov models (HMMs), the complexity is the same as that specialized for HMMs. As an example to eliminate restrictions of existing approaches, we utilize our algorithm to give diagnoses for failure in a logic circuit involving stochastic error gates.

12. Exploiting the High Predictive Power of Multi-class Subgroups
    Tarek Abudawood (University of Bristol) and Peter Flach (University of Bristol)

    Subgroup discovery aims at finding subsets of a population whose class distribution is significantly different from the overall distribution. A number of multi-class subgroup discovery methods has been previously investigated, proposed and implemented in the CN2-MSD system. When a decision tree learner was applied using the induced subgroups as features, it led to the construction of accurate and compact predictive models, demonstrating the usefulness of the subgroups. In this paper we show that, given a significant, sufficient and diverse set of subgroups, no further learning phase is required to build a good predictive model. Our systematic study bridges the gap between rule learning and decision tree modelling by proposing a method which uses the training information associated with the subgroups to form a simple tree-based probability estimator and ranker, RankFree-MSD, without the need for an additional learning phase. Furthermore, we propose an efficient subgroup pruning algorithm, RankFree-Pruning, that prunes unimportant subgroups from the subgroup tree in order to reduce the number of subgroups and the size of the tree without decreasing predictive performance. Despite the simplicity of our approach we experimentally show that its predictive performance in general is comparable to other decision tree and rule learners over 10 multi-class UCI data sets.

13. Generative Models of Information Diffusion with Asynchronous Time-delay

    Kazumi Saito (University of Shizuoka), Masahiro Kimura (Ryukoku University), Kouzou Ohara (Aoyama Gakuin University), and Hiroshi Motoda (Osaka University)

    We address the problem of formalizing an information diffusion process on a social network as a generative model in the machine learning framework so that we can learn model parameters from the observation. Time delay plays an important role in formulating the likelihood function as well as for the analyses of information diffusion. We identified that there are two different types of time delay: link delay and node delay. The former corresponds to the delay associated with information propagation, and the latter corresponds to the delay due to human action. We further identified that there are two distinctions of the way the activation from the multiple parents is updated: non-override and override. The former sticks to the initial activation and the latter can decide to update the time to activate multiple times. We formulated the likelihood function of the well known diffusion models: independent cascade and linear threshold, both enhanced with asynchronous time delay distinguishing the difference in two types of delay and two types of update scheme. Simulation using four real world networks reveals that there are differences in the spread of information diffusion and they strongly depend on the choice of the parameter values and the denseness of the network.

14. Accurate Ensembles for Data Streams: Combining Restricted Hoeffding Trees using Stacking
Albert Bifet (University of Waikato), Eibe Frank (University of Waikato), Geoffrey Holmes (University of Waikato), and Bernhard Pfahringer (University of Waikato)
The success of simple methods for classification shows that is is often not necessary to model complex attribute interactions to obtain good classification accuracy on practical problems. In this paper, we propose to exploit this phenomenon in the data stream context by building an ensemble of Hoeffding trees that are each limited to a small subset of attributes. In this way, each tree is restricted to model interactions between attributes in its corresponding subset. Because it is not known a priori which attribute subsets are relevant for prediction, we build exhaustive ensembles that consider all possible attribute subsets of a given size. As the resulting Hoeffding trees are not all equally important, we weigh them in a suitable manner to obtain accurate classifications. This is done by combining the log-odds of their probability estimates using sigmoid perceptrons, with one perceptron per class. We propose a mechanism for setting the perceptrons' learning rate using the ADWIN change detection method for data streams, and also use ADWIN to reset ensemble members (i.e. Hoeffding trees) when they no longer perform well. Our experiments show that the resulting ensemble classifier outperforms bagging for data streams in terms of accuracy when both are used in conjunction with adaptive naive Bayes Hoeffding trees, at the expense of runtime and memory consumption.

15. Mining Recurring Concept Drifts with Limited Labeled Streaming Data
Peipei Li, (Hefei University of Technology), Xindong Wu (University of Vermont), and Xuegang Hu (Hefei University of Technology)
Tracking recurring concept drifts is a significant issue for machine learning and data mining that frequently appears in real world stream classification problems. It is a challenge for many streaming classification algorithms to learn recurring concepts in a data stream envi-

ronment with unlabeled data, and this challenge has received little attention from the research community. Motivated by this challenge, this paper focuses on the problem of recurring contexts in streaming environments with limited labeled data. We propose a Semi-supervised classification algorithm for data streams with REcurring concept Drifts and Limited LAbeled data, called REDLLA, in which, a decision tree is adopted as the classification model. When growing a tree, a clustering algorithm based on k-Means is installed to produce concept clusters and unlabeled data are labeled at leaves. In view of deviations between history and new concept clusters, potential concept drifts are distinguished and recurring concepts are maintained. Extensive studies on both synthetic and real-world data confirm the advantages of our REDLLA algorithm over two state-of-the-art online classification algorithms of CVFDT and CDRDT and several known online semi-supervised algorithms, even in the case with more than 90% unlabeled data.

16. Hierarchical Convex NMF for Clustering Massive Data
    Kristian Kersting (Fraunhofer IAIS and University of Bonn), Mirwaes Wahabzada (Fraunhofer IAIS), Christian Thurau (Fraunhofer IAIS), and Christian Bauckhage (Fraunhofer IAIS)

    We present an extension of convex-hull non-negative matrix factorization (CH-NMF) which was recently proposed as a large scale variant of convex non-negative matrix factorization or Archetypal Analysis. CH-NMF factorizes a non-negative data matrix V into two non-negative matrix factors V = WH such that the columns of W are convex combinations of certain data points so that they are readily interpretable to data analysts. There is, however, no free lunch: imposing convexity constraints on W typically prevents adaptation to intrinsic, low dimensional structures in the data. Alas, in cases where the data is distributed in a non-convex manner or consists of mixtures of lower dimensional convex distributions, the cluster representatives obtained from CH-NMF will be less meaningful. In this paper, we present a hierarchical CH-NMF that automatically adapts to internal structures of a dataset, hence it yields meaningful and interpretable clusters for non-convex datasets. This is also confirmed by our extensive evaluation on DBLP publication records of 760,000 authors, 4,000,000 images harvested from the web, and 150,000,000 votes on World of Warcraft guilds.

17. Multi-task Learning for Recommender System
    Xia Ning (University of Minnesota) and George Karypis (University of Minnesota)

    This paper focuses on exploring personalized multi-task learning approaches for collaborative filtering towards the goal of improving the prediction performance of rating prediction systems. These methods first specifically identify a set of users that are closely related to the user under consideration (i.e., active user), and then learn multiple rating prediction models simultaneously, one for the active user and one for each of the related users. Such learning for multiple models (tasks) in parallel is implemented by representing all learning instances (users and items) using a coupled user-item representation, and within error-insensitive Support Vector Regression (e-SVR) framework applying multi-task kernel tricks. A comprehensive set of experiments shows that multi-task learning approaches lead to significant performance improvement over conventional alternatives.

18. Adaptive Step-size Policy Gradients with Average Reward Metric
    Takamitsu Matsubara (NAIST/ATR), Tetsuro Morimura (IBM Research), and Jun Morimoto (ATR)

    In this paper, we propose a novel adaptive step-size approach for policy gradient reinforcement learning. A new metric is defined for policy gradients that measures the effect of changes on average reward with respect to the policy parameters. Since the metric directly measures the effects on the average reward, the resulting policy gradient learning employs an adaptive step-size strategy that can effectively avoid falling into a stagnant phase from the complex structure of the average reward function with respect to the policy parameters. Two algorithms are derived with the metric as variants of ordinary and natural policy gradients. Their properties are compared with previously proposed policy gradients through numerical experiments with simple, but non-trivial, 3-state Markov Decision Processes (MDPs). We also show performance improvements over previous methods in on-line learning with more challenging 20-state MDPs.

19. Finite-sample Analysis of Bellman Residual Minimization
Odalric-Ambrym Maillard (INRIA Lille Nord-Europe), Remi Munos (INRIA Lille Nord-Europe), Alessandro Lazaric (INRIA Lille Nord-Europe), and Mohammad Ghavamzadeh (INRIA Lille Nord-Europe)

    We consider the Bellman residual minimization approach for solving discounted Markov decision problems, where we assume that a generative model of the dynamics and rewards is available. At each policy iteration step, an approximation of the value function for the current policy is obtained by minimizing an empirical Bellman residual defined on a set of n states drawn i.i.d. from a distribution μ, the immediate rewards, and the next states sampled from the model. Our main result is a generalization bound for the Bellman residual in linear approximation spaces. In particular, we prove that the empirical Bellman residual approaches the true (quadratic) Bellman residual in μ-norm with a rate of order $O(1/\sqrt{n})$. This result implies that minimizing the empirical residual is indeed a sound approach for the minimization of the true Bellman residual which guarantees a good approximation of the value function for each policy. Finally, we derive performance bounds for the resulting approximate policy iteration algorithm in terms of the number of samples n and a measure of how well the function space is able to approximate the sequence of value functions.

20. A Study of Approximate Inference in Probabilistic Relational Models
Fabian Kaelin (McGill) and Doina Precup (McGill)

    We tackle the problem of approximate inference in Probabilistic Relational Models (PRMs) and propose the Lazy Aggregation Block Gibbs (LABG) algorithm. The LABG algorithm makes use of the inherent relational structure of the ground Bayesian network corresponding to a PRM. We evaluate our approach on artificial and real data and show that it scales well with the size of the data set.

21. Conceptual Imitation Learning: An Application to Human-robot Interaction
Hossein Hajimirsadeghi (University of Tehran), Majid Nili Ahmadabadi (University of Tehran), Mostafa Ajallooeian (University of Tehran), Babak Araabi (University of Tehran), and Hadi Moradi (University

of Tehran)

In general, imitation is imprecisely used to address different levels of social learning from high level knowledge transfer to low level regeneration of motor commands. However, true imitation is based on abstraction and conceptualization. This paper presents a conceptual approach for imitation learning using feedback cues and interactive training to abstract spatio-temporal demonstrations based on their perceptual and functional characteristics. Abstraction, concept acquisition, and self-organization of proto-symbols are performed through an incremental and gradual learning algorithm. In this algorithm, Hidden Markov Models (HMMs) are used to abstract perceptually similar demonstrations. However, abstract (relational) concepts emerge as a collection of HMMs irregularly scattered in the perceptual space. Performance of the proposed algorithm is evaluated in a human-robot interaction task of imitating signs produced by hand movements. Experimental results show efficiency of our model for concept extraction, symbol emergence, motion pattern recognition, and regeneration.

Accepted Posters

1. Utilizing Fuzzy-SVM and a Subject Database to Reduce the Calibration Time of P300-based BCI
   Sercan Taha Ahi (Tokyo Institute of Technology), Natsue Yoshimura (Tokyo Institute of Technology), Hiroyuki Kambara (Tokyo Institute of Technology), and Yasuharu Koike (Tokyo Institute of Technology)

2. Feature Selection for Reinforcement Learning: Evaluating Implicit State-reward Dependency via Conditional Mutual Information
   Hirotaka Hachiya (Tokyo Institute of Technology) and Masashi Sugiyama (Tokyo Institute of Technology)

3. Dependence Minimizing Regression with Model Selection for Non-linear Causal Inference under Non-Gaussian Noise
   Makoto Yamada (Tokyo Institute of Technology) and Masashi Sugiyama (Tokyo Institute of Technology)

4. Joint Unsupervised Learning of Parallel Sequence Alignment and Segmentation
   Mark Fishel (University of Tartu)

5. Multi-class Subgroup Discovery
   Tarek Abudawood (University of Bristol) and Peter Flach (University of Bristol)

6. A Comparison of CNF with CRF in Named Entity Recognition Task
   Kei Uchiumi (Yahoo Japan Corporation), Keigo Machinaga (Yahoo Japan Corporation), Toshiyuki Maezawa (Yahoo Japan Corporation), and Toshinori Satou (Yahoo Japan Corporation)

7. Multiscale-bagging with Applications to Classification
   Masayoshi Aoki (Tokyo Institute of Technology), Takafumi Kanamori (Nagoya University), and Hidetoshi Shimodaira (Tokyo Institute of Technology)

8. Contrasting Correlations by an Efficient Double-clique Search Method
   Aixiang Li (Hokkaido University) and Makoto Haraguchi (Hokkaido University)

9. Model-induced Regularization
Shinichi Nakajima (Nikon Corporation) and Masashi Sugiyama (Tokyo Institute of Technology)

10. Slice Sampling on Chinese Restaurant Process
Takaki Makino (University of Tokyo)

11. Interactive Behavior Adaptation through Dialogue Based on Bayesian Network
Saifuddin Md. Tareeq (The Graduate University for Advanced Studies) and Tetsunari Inamura (National Institute of Informatics)

12. Maximum Volume Clustering
Gang Niu (Nanjing University), Bo Dai (Chinese Academy of Science), Lin Shang (Nanjing University), and Masashi Sugiyama (Tokyo Institute of Technology)

13. Using Conditional Random Fields to Validate Observations in a 4W1H Paradigm
Leon F. Palafox (University of Tokyo), Laszlo A. Jeni (University of Tokyo), and Hideki Hashimoto (University of Tokyo)

14. Multiscale Bagging with Applications to Classification and Active Learning
Hidetoshi Shimodaira (Tokyo Institute of Technology), Takafumi Kanamori (Nagoya University), Masayoshi Aoki (Tokyo Institute of Technology), and Kouta Mine (Tokyo Institute of Technology)

15. Adjustment for Multiple Hypotheses Testing in Comparative Classification Studies
Daniel Berrar (Tokyo Institute of Technology)

16. Inference in Latent Conditional Models: The Computational Complexity Analysis and a Comparative Study of Solutions
Xu Sun (University of Tokyo), Hisashi Kashima (University of Tokyo), and Takuya Matsuzaki (University of Tokyo)

17. Improving Graph-based Semi-supervised Learning by Feature Space Transformation
Yu-Shi Lin (Academia Sinica and National Taiwan University) and

Chun-Nan Hsu (Academia Sinica and University of Southern California)

18. Image Annotation via Multi-instance Learning with Pyramid Graph Kernel
    Zhi Nie (Tsinghua University), Guiguang Ding (Tsinghua University), and Chunping Li (Tsinghua University)

19. Proximity in Large Bipartite Graphs with Unsupervised Auxiliary Information
    Rudy Raymond (IBM Research - Tokyo), Yuta Tsuboi (IBM Research - Tokyo), Hisashi Kashima (The University of Tokyo), and Issei Sato (The University of Tokyo)

Organizers

General Chair

- Takashi Washio (Osaka University, Japan)

Program Committee Co-chairs

- Masashi Sugiyama (Tokyo Institute of Technology, Japan)
- Qiang Yang (Hong Kong University of Science and Technology, China)

Local Arrangement Chair

- Hirotaka Hachiya (Tokyo Institute of Technology, Japan)

ACML Steering Committee Co-chairs

- Hiroshi Motoda (AFOSR/AOARD and Osaka University, Japan)
- Zhi-Hua Zhou (Nanjing University, China)

ACML Steering Committee

- Thomas G. Dietterich (Oregon State University, USA)
- Tu Bao Ho (JAIST, Japan)
- Bernhard Pfahringer (Waikato University, New Zealand)
- Takashi Washio (Osaka University, Japan)
- Geoff Webb (Monash University, Australia)
- Qiang Yang (Hong Kong University of Science and Technology, China)

Senior Program Committee

- Sung-Bae Cho (Yonsei University, Korea)
- Wei Fan (IBM T.J.Watson Research, USA)
- Eibe Frank (University of Waikato, New Zealand)
- Joaõ Gama (University of Porto, Portugal)
- James Kwok (Hong Kong University of Science and Technology, China)
- Hang Li (Microsoft Research Asia, China)
- Charles Ling (University of Western Ontario, Canada)
- Chih-Jen Lin (National Taiwan University, Taiwan)
- Huan Liu (Arizona State University, USA)
- Yuji Matsumoto (Nara Institute of Science and Technology, Japan)
- Katharina Morik (University of Dortmund, Germany)
- Klaus-Robert Müller (Technical University Berlin, Germany)
- Kai Ming Ting (Monash University, Australia)
- Koji Tsuda (National Institute of Advanced Industrial Science and Technology, Japan)
- Jean-Philippe Vert (Mines ParisTech, France)
- Sethu Vijayakumar (University of Edinburgh, UK)
- Limsoon Wong (National University of Singapore, Singapore)
- Kenji Yamanishi (University of Tokyo, Japan)
- Changshui Zhang (Tsinghua University, China)

Program Committee

- Shin Ando (Gunma University)
- Peter Andreae (Victoria University of Wellington)
- Annalisa Appice (University of Bari)
- Marta Arias (Universitat Politecnica de Catalunya)
- Antonio Bahamonde (Universidad de Oviedo at Gijon)
- Daniel Berrar (Tokyo Institute of Technology)
- Gilles Bisson (CNRS)
- Christian Bockermann (Technical University of Dortmund)
- Gilles Blanchard (Weierstrass Institute for Applied Analysis and Stochastics)

- Wray Buntine (NICTA)
- Tru Hoang Cao (Ho Chi Minh City University of Technolgoy)
- Songcan Chen (Nanjing University of Aeronautics & Astronautics)
- Xilin Chen (Institute of Computing Technology, Chinese Academy of Sciences)
- Li Cheng (TTI-Chicago)
- Hong Cheng (Chinese University of Hong Kong)
- Jian Cheng (Institute of Automation, Chinese Academy of Sciences)
- Mingmin Chi (Fudan University)
- Yun Chi (NEC Laboratories America)
- Marco Cuturi (Princeton University)
- Alfredo Cuzzocrea (ICAR-CNR and University of Calabria)
- Andre de Carvalho (Universidade de Sao Paulo)
- Jos del Campo-Avila (University of Malaga)
- Fazel Famili (Institute For Information Technology, National Research Council of Canada)
- Dragan Gamberger (Rudjer Boskovic Institute)
- Yang Gao (Nanjing University)
- Mao-Zu Guo (Harbin Institute of Technology)
- Yuhong Guo (Temple University)
- Hirotaka Hachiya (Tokyo Institute of Technology)
- Makoto Haraguchi (Hokkaido University)
- Kohei Hatano (Kyushu University)
- Xiaofei He (Zhejiang University)
- Kouichi Hirata (Kyushu Institute of Technology)
- Xian-Sheng Hua (Microsoft Research Asia)
- Tsuyoshi Ide (IBM Research, Tokyo Research Laboratory)
- Akihiro Inokuchi (Osaka University)
- Tomoharu Iwata (NTT)
- Kietz Jorg-Uwe (Univ. Zurich)
- Sanjay Jain (National University of Singapore)
- Furnkranz Johannes (TU Darmstadt)
- Hisashi Kashima (University of Tokyo)

- Tsuyoshi Kato (Ochanomizu University)
- Yanai Keiji (University of Electro-Communications)
- Yamazaki Keisuke (Tokyo Institute of Technology)
- Kristian Kersting (Fraunhofer IAIS and University of Bonn)
- Akisato Kimura (NTT)
- Manabu Kimura (Tokyo Institute of Technology)
- Stefan Kramer (Technische Universitat Munchen)
- Nicole Kramer (Weierstrass Institute for Applied Analysis and Stochastics)
- Rui Kuang (University of Minnesota)
- Tetsuji Kuboyama (Gakushuin University)
- Vincent Lemaire (Orange Labs R&D)
- Chun-hung Li (Hong Kong Baptist University)
- Gang Li (Deakin University)
- Guo-Zheng Li (Tongji University)
- Jinyan Li (Nanyang Technological University)
- Ming Li (Nanjing University)
- Tao Li (Florida International University)
- Xiaoli Li (Infocomm)
- Xue Li (The University of Queensland)
- Xuelong Li (University of London)
- Zhouchen Lin (Microsoft Research Asia)
- Hiroshi Mamitsuka (Kyoto University)
- Michael May (Fraunhofer IAIS)
- Shin-ichi Minato (Hokkaido University)
- Dunja Mladenic (Jozef Stefan Institute)
- Daichi Mochihashi (NTT)
- Tetsuro Morimura (IBM Research, Tokyo Research Laboratory)
- Hiroshi Nakagawa (University of Tokyo)
- Shinichi Nakajima (Nikon Corporation)
- See-Kiong Ng (Institute for Infocomm Research, A*STAR)
- Shigeyuki Oba (Kyoto University)
- Tomonobu Ozaki (Kobe University)

- Junfeng Pan (Google)
- Seong-Bae Park (Kyungpook National University)
- Yvan Saeys (Ghent University)
- Jun Sakuma (Tsukuba University)
- Taisuke Sato (Tokyo Institute of Technology)
- Nazha Selmaoui (PPME/ERIM)
- Jun Sese (Ochanomizu University)
- Rudy Setiono (National University of Singapore)
- Dou Shen (Microsoft adCenter Labs)
- Masashi Shimbo (NAIST)
- Shohei Shimizu (Osaka University)
- Hyunjung (Helen) Shin (Ajou University)
- Le Song (Carnegie Mellon University)
- Eduardo Spinosa (Federal University of Parana)
- Jian-Tao Sun (Microsoft Research Asia)
- Taiji Suzuki (University of Tokyo)
- Ichiro Takeuchi (Nagoya Institute of Technology)
- Ah-Hwee Tan (Nanyang Technological University)
- Dacheng Tao (Nanyang Technological University)
- Alexandre Termier (LIG, University of Grenoble)
- Ryota Tomioka (University of Tokyo)
- Ivor Tsang (Nanyang Technological University)
- Vincent S. Tseng (National Cheng Kung University)
- Yuta Tsuboi (IBM Research, Tokyo Research Laboratory)
- Takeaki Uno (National Institute of Informatics (NII) of JAPAN)
- Fei Wang (Florida International University)
- Hansheng Wang (Peking University)
- Jun Wang (Syracuse University)
- Lei Wang (The Australian National University)
- Liwei Wang (Peking University)
- Shijun Wang (U.S. National Institutes of Health)
- Xizhao Wang (Hebei University)
- Gui-Rong Xue (Shanghai Jiaotong University)

- Takehisa Yairi (University of Tokyo)
- Makoto Yamada (Tokyo Institute of Technology)
- Shuicheng Yan (NUS/ECE)
- Jun Yan (Microsoft Research Asia)
- Xifeng Yan (University of California at Santa Barbara)
- Jieping Ye (Arizon State University)
- Jie (Jessie) Yin (CSIRO ICT Centre)
- Kai Yu (NEC Lab America)
- Lei Yu (Binghamton University)
- Shipeng Yu (Siemens)
- Filip Zelezny (Czech Technical University in Prague)
- Bo Zhang (Tsinghua University)
- Dell Zhang (Birkbeck, University of London)
- Daoqiang Zhang (Nanjing University of Aeronautics and Astronautics)
- Harry Zhang (University of New Brunswick)
- Junping Zhang (Fudan University)
- Kai Zhang (Lawrence Berkeley National Lab)
- Kun Zhang (Helsinki University of Technology)
- Min-Ling Zhang (Hohai University)
- Zhongfei (Mark) Zhang (SUNY Binghampton)
- Yan Zhou (University of South Alabama)
- Xingquan Zhu (Florida Atlantic University)

Memo

Memo