Poster Abstract P01:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

Utilizing Fuzzy-SVM and a Subject Database to Reduce the Calibration Time of P300-based BCI

Sercan Taha Ahi[1], Natsue Yoshimura[2], Hiroyuki Kambara[2], and Yasuharu Koike[3]

[1]Department of Computational Intelligence and System Science
[2]Precision and Intelligence Laboratory
[3]Solution Science Research Laboratory
Tokyo Institute of Technology,
4259-R2-15, Nagatsuta, Midori-ku, Yokohama 226-8503 JAPAN
taha.ahi@hi.pi.titech.ac.jp,yoshimura@cns.pi.titech.ac.jp
hkambara@hi.pi.titech.ac.jp,koike@pi.titech.ac.jp

Abstract. Current Brain-Computer Interfaces (BCI) suffer the requirement of a subject-specific calibration process due to variations in EEG responses across different subjects. Additionally, the duration of the calibration process should be long enough to sufficiently sample high dimensional feature spaces. In this study, we proposed a method based on Fuzzy Support Vector Machines (Fuzzy-SVM) and a database of training samples from several subjects to address both issues for P300-based BCI. To validate the proposed approach, we conducted P300 speller experiments on 18 subjects and formed a subject-database using the leave-one-out approach. Fuzzy-SVM is an extension to the traditional SVM in which a different weight is assigned to every slack variable. We assigned the same weight to all the slack variables coming from a specific subject in the database. The weight of a subject in the database set to be proportional to the accuracy obtained by a standard SVM which is trained using only samples from the corresponding subject and tested with samples of the test-subject. With the proposed approach, we achieved to obtain an average accuracy of 80% with only 4 training letters. Conventional subject-specific calibration approach, on the other hand, needed 12 training letters to provide the same performance.

Keywords: Brain-Computer Interfaces, P300, EEG, Subject-Database, Fuzzy Support Vector Machines

Poster Abstract P02:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

Feature Selection for Reinforcement Learning: Evaluating Implicit State-Reward Dependency via Conditional Mutual Information

Hirotaka Hachiya and Masashi Sugiyama
Tokyo Institute of Technology, Tokyo, 152-8552, Japan.

Abstract

Model-free reinforcement learning (RL) is a machine learning approach to decision making in unknown environments. However, real-world RL tasks often involve high-dimensional state spaces, and then standard RL methods do not perform well. In this work, we propose a new feature selection framework for coping with high dimensionality. Our proposed framework adopts *conditional mutual information* between return and state-feature sequences as a feature selection criterion, allowing the evaluation of implicit state-reward dependency. The conditional mutual information is approximated by a least-squares method, which results in a computationally efficient feature selection procedure. The usefulness of the proposed method is demonstrated on grid-world navigation problems.

1 Introduction

Optimal decision making in unknown environment is a challenging task in the machine learning community. *Reinforcement learning* (RL) is a popular framework for this purpose, where a *policy* (the decision rule of an agent) is determined so that *return* (the sum of discounted rewards the agent will receive) is maximized. However, when the dimensionality of the state space is high, existing RL approaches tends to perform poorly. Unfortunately, this critically limits the range of applicability of RL in practice since real-world RL tasks (e.g., robot control) often involve high-dimensional state spaces. To cope with high dimensionality of the state space, choosing a subset of relevant features from the high-dimensional state variables, i.e., *feature selection*, is highly useful.

In this work, we introduce a new framework of filter-type feature selection for RL where we evaluate the independence between return and state-feature sequences using the conditional *mutual information* (MI) [1]. In order to efficiently approximate the conditional MI from samples, we utilize a least-squares MI estimator which was proved to possess the optimal convergence rate [2].

2 Feature Selection via Conditional Mutual Information

In this section, we briefly describe our proposed feature selection method.

Let η_n and $\boldsymbol{s}_n = (s_n^{(1)}, s_n^{(2)}, \ldots, s_n^{(v)})$ be the return and the v dimensional state features at the n-th time step. For u ($\leq v$) being the number of features we want to select, our goal is to find a 'subset' $\boldsymbol{z}_n = (z_n^{(1)}, z_n^{(2)}, \ldots, z_n^{(u)})^\top$ of the state features $\boldsymbol{s}_n$ such that

$$\eta_n \perp \boldsymbol{s}_n \mid \boldsymbol{z}_n, \quad \forall n = 1, 2, \ldots, N. \tag{1}$$

This means that, for all time steps, the return η_n is conditionally independent of the entire state features $\boldsymbol{s}_n$ given the subset $\boldsymbol{z}_n$.

We propose to use *conditional MI* $I(\eta; \mathrm{z}|\mathrm{n})$ as our feature selection criterion, which is defined as the average of MI $I(\eta_n; \mathrm{z}_n)$ over time steps $n = 1, 2, \ldots, N$ [1]:

$$I(\eta; \mathrm{z}|\mathrm{n}) = \frac{1}{N} \sum_{n=1}^{N} I(\eta_n; \mathrm{z}_n).$$

The conditional MI between returns and state features can be seen as a measure of dependency between returns and state-feature sequences. The rationale behind the use of conditional MI for feature selection relies on the following lemma(its proof is omitted due to the limited space):

Lemma 1

$$I(\eta; \mathrm{s}|\mathrm{n}) - I(\eta; \mathrm{z}|\mathrm{n}) = \frac{1}{N} \sum_{n=1}^{N} \iint \frac{p(\eta, \boldsymbol{z}|n)^2}{p(\eta|n)p(\boldsymbol{z}|n)^2} \times \left(\frac{p(\eta, \boldsymbol{s}|\boldsymbol{z}, n)}{p(\boldsymbol{s}|\boldsymbol{z}, n)p(\eta|\boldsymbol{z}, n)} - 1 \right)^2 p(\boldsymbol{s}|n) \mathrm{d}\boldsymbol{s} \mathrm{d}\eta \geq 0.$$

This lemma implies that $I(\eta; \mathrm{s}|\mathrm{n}) \geq I(\eta; \mathrm{z}|\mathrm{n})$ and the equality holds if and only if

$$p(\eta, \boldsymbol{s}|\boldsymbol{z}, n) = p(\eta|\boldsymbol{z}, n)p(\boldsymbol{s}|\boldsymbol{z}, n), \ \forall n = 1, 2, \ldots, N.$$

This is equivalent to Eq.(1), and thus Eq.(1) can be attained by maximizing $I(\eta; \mathrm{z}|\mathrm{n})$ with respect to $\boldsymbol{z}$.

In this work, we employ *forward-selection* [3] algorithm to find the subset $\boldsymbol{z}$ which maximizing the conditional MI $I(\eta; \mathrm{z}|\mathrm{n})$ while utilizing *least-squares MI* [2] to approximate $I(\eta; \mathrm{z}|\mathrm{n})$.

References

[1] MacKay, D.J.C.: Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge, UK (2003)

[2] Suzuki, T., Sugiyama, M., Kanamori, T., Sese, J.: Mutual information estimation reveals global associations between stimuli and biological processes. BMC Bioinformatics **10**(1) (2009) S52

[3] Song, L., Smola, A., Gretton, A., Borgwardt, K., Bedo, J.: Supervised feature selection via dependence estimation. In: Proceedings of the 24th International Conference on Machine Learning. (2007) 823–830

Poster Abstract P03:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

Dependence Minimizing Regression with Model Selection for Non-Linear Causal Inference under Non-Gaussian Noise *†

Makoto Yamada† and **Masashi Sugiyama**†‡
†Department of Computer Science, Tokyo Institute of Technology
‡Japan Science and Technology Agency
{yamada@sg. sugi@}cs.titech.ac.jp

Introduction

The discovery of non-linear causal relationship under additive non-Gaussian noise models has attracted considerable attention recently because of their high flexibility. In this poster, we propose a novel causal inference algorithm called *least-squares independence regression* (LSIR). LSIR learns the additive noise model through minimization of an estimator of the *squared-loss mutual information* between inputs and residuals. A notable advantage of LSIR over existing approaches is that tuning parameters such as the kernel width and the regularization parameter can be naturally optimized by cross-validation, allowing us to avoid overfitting in a data-dependent fashion. Through experiments with real-world datasets, we show that LSIR compares favorably with the state-of-the-art causal inference method.

Dependence Minimizing Regression by LSIR

Suppose random variables $X \in \mathbb{R}$ and $Y \in \mathbb{R}$ are connected by the following additive noise model (Hoyer et al. 2009):

$$Y = f(X) + E,$$

where $f(\cdot) : \mathbb{R} \to \mathbb{R}$ is some non-linear function and $E \in \mathbb{R}$ is a zero-mean random variable independent of X. The goal of dependence minimizing regression is, from i.i.d. paired samples $\{(x_i, y_i)\}_{i=1}^n$, to obtain a function $\widehat{f}(\cdot)$ such that input X and estimated additive noise $\widehat{E} = Y - \widehat{f}(X)$ are independent.

Let us employ a linear model for dependence minimizing regression $f_{\boldsymbol{\beta}}(x) = \sum_{l=1}^m \beta_l \psi_l(x) = \boldsymbol{\beta}^\top \boldsymbol{\psi}(x)$, where m is the number of basis functions, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^\top$ are regression parameters, $^\top$ denotes the transpose, and $\boldsymbol{\psi}(x) = (\psi_1(x), \ldots, \psi_m(x))^\top$ are Gaussian basis functions.

In dependence minimization regression, the regression parameter $\boldsymbol{\beta}^*$ may be learned as

$$\boldsymbol{\beta}^* = \mathop{\mathrm{argmin}}_{\boldsymbol{\beta}} \left[\mathrm{SMI}(X, \widehat{E}) + \frac{\gamma}{2} \boldsymbol{\beta}^\top \boldsymbol{\beta} \right],$$

where $\gamma \geq 0$ is the regularization parameter for avoiding overfitting, and $\mathrm{SMI}(X, \widehat{E})$ is the *squared-loss mutual information* (SMI) between X and $\widehat{E}$:

$$\mathrm{SMI}(X, \widehat{E}) = \frac{1}{2} \iint \frac{p(x, \widehat{e})}{p(x) p(\widehat{e})} p(x, \widehat{e}) \mathrm{d}x \mathrm{d}\widehat{e} - \frac{1}{2}.$$

SMI cannot be directly computed since it contains unknown densities $p(x, \widehat{e})$, $p(x)$, and $p(\widehat{e})$. Thus, we use *least-squares mutual information* (LSMI) (Suzuki et al. 2009) to estimate an empirical SMI. A key idea of LSMI is to directly estimate the *density ratio*:

$$w(x, \widehat{e}) = \frac{p(x, \widehat{e})}{p(x) p(\widehat{e})},$$

without going through density estimation of $p(x, \widehat{e})$, $p(x)$, and $p(\widehat{e})$.

Given a density ratio estimator $\widehat{w} = w_{\widehat{\boldsymbol{\alpha}}}$ estimated by LSMI, SMI can be simply approximated as

$$\widehat{\mathrm{SMI}}(X, \widehat{E}) = \frac{1}{2n} \sum_{i=1}^n \widehat{w}(x_i, \widehat{e}_i) - \frac{1}{2}. \quad (1)$$

Finally, we learn the regression parameter $\boldsymbol{\beta}^*$ so that $\widehat{\mathrm{SMI}}(X, \widehat{E})$ is minimized.

We call this SMI based dependence minimizing regression as *least-squares independence regression* (LSIR). A notable advantage of LSIR over existing approaches is that tuning parameters such as the kernel width and the regularization parameter can be naturally optimized by cross-validation, allowing us to avoid overfitting in a data-dependent fashion.

Causal Direction Inference

Our final goal is, given i.i.d. paired samples $\{(x_i, y_i)\}_{i=1}^n$, to determine whether X causes Y or vice versa. To this end, we test whether the causal model $Y = f_Y(X) + E_Y$ or the alternative model $X = f_X(Y) + E_X$ fits the data well, where the goodness of fit is measured by independence between inputs and residuals (i.e., estimated noise). In this poster, independence of inputs and residuals are decided by the *permutation test* (Efron and Tibshirani 1993) with the use of LSMI.

References

Efron, B., and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall.

Hoyer, P. O.; Janzing, D.; Mooij, J. M.; Peters, J.; and Schölkopf, B. 2009. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21 (NIPS2008)*, 689–696. Cambridge, MA: MIT Press.

Suzuki, T.; Sugiyama, M.; Kanamori, T.; and Sese, J. 2009. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics* 10(S52).

*http://sugiyama-www.cs.titech.ac.jp/ sugi/2010/AAAI2010.pdf.

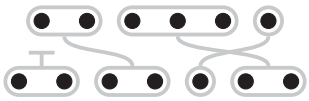
†This work was supported by SCAT, AOARD, and the JST PRESTO program.

Poster Abstract P04:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

Joint Unsupervised Learning of Parallel Sequence Alignment and Segmentation

Mark Fishel
Institute of Computer Science
University of Tartu, Estonia
fishel@ut.ee

The focus of our work is unsupervised learning the optimal segmentation of two symbol sequences and alignment of the resulting segments to optimize a joint objective, where a sequence segmentation is a partition of the ordered set of its symbols, where each part must only contain consequent symbols. The objective function is defined in terms of the model parameters, which are the individual weights of single segments and aligned segment pairs. The following is an example of two segmented and aligned sequences:



Here the first group of the bottom sequence is unaligned.

We propose two approaches to learning the parameters of the model. The first one is based on the Expectation-Maximization algorithm, where the segmentations and the alignment both constitute the hidden variable Z. The maximization step thus involves summing the products of the probabilities of all segments of each segmentation, which in our case can be accomplished with a dynamic programming-based algorithm with a quadratic time complexity. To stabilize the iterative search and initialize the parameters the first few iterations are made only for segmentations, before proceeding with the full search.

The second approach is based on Eric Brill's transformation-based learning (TBL) with the error function to minimize is defined via the minimum description length principle. The search space is reduced with random subsampling of the possible search directions, commonly employed with TBL. As soon as the search is converged, the actual parameters are estimated with the maximum-likelihood principle from the segmented and aligned training set.

In addition to learning it is necessary to solve the task of finding the optimal segmentations and alignments for unseen samples, once the parameters have been learned. We combine Viterbi search with maximum a-posteriori decoding to avoid the exponential time complexity of pure Viterbi search. Decoding is guided with a parameter $\alpha \in [0, 1]$, which allows finding a suitable compromise between quality and decoding time.

The main domain of applying the proposed model is statistical machine translation, where it can be put to many uses. It can be applied to languages with no explicit word boundaries, like Chinese and Japanese, and also morphologically rich languages, like Finnish or Turkish, to integrate word alignment with segmenting the sentence into words or highly inflectional word forms into morphemes to reduce data sparsity. Other approaches have shown that linguistic and monolingual segmentation is not necessary optimal, depending on the source and target language of translation, which makes our essentially bilingual approach promising.

At the same time one of the state-of-the-art approaches to machine translation is phrase-based statistical translation, which involves finding the correspondence of the word sequences between sentences in the training set. Our approach can be substituted for the currently used heuristic learning model of this kind of translation.

The main question is currently whether the described approach works for the suggested applications. Initial experiments show that the approach works well on "toy" training/testing data. Work in progress is mainly focused on efficiently implementing the described models and testing the approach in full scale machine translation.

Poster Abstract P05:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

Multi-class Subgroup Discovery

Tarek Abudawood (Dawood@cs.bris.ac.uk) and
Peter Flach (Peter.Flach@bristol.ac.uk)

Intelligent Systems Laboratory
University of Bristol
United Kingdom

Rule induction is a common form of machine learning and data mining often used in classification and association rule learning. Classification rule learning is a predictive task aimed at constructing a set of rules, based on training examples and their observed features, to predict the class of unseen future examples. Association rule learning, on the other hand, is a form of descriptive induction aimed at the discovery of individual rules that express interesting patterns in data.

In classification rule learning a target concept is pre-defined and so the search heuristic is usually some form of accuracy. On the other hand, in descriptive rule learning no target concept is given and the heuristic function evaluates measures of interestingness and unusualness in the data, e.g. support and confidence. Subgroup discovery can be seen as being halfway between predictive and descriptive rule learning, as there is a target concept but the goal of subgroup discovery is not necessarily to achieve high accuracy. Rather, the target concept helps us to achieve a trade-off between accuracy and interestingness. Subgroup discovery aims at finding subsets of a population whose class distribution is significantly different from the overall distribution.

Subgroup discovery has previously predominantly been investigated in a two-class context such that each discovered subgroup reflects an interesting phenomenon occurs in a single class and the discovery is guided by a two-class heuristic measures. Recently subgroup discovery has been investigated in a multi-class context where the discovered subgroups express interesting phenomena which may occur not only in a single class but rather in multiple classes where multi-class heuristics where studied and used to derive such subgroups.

We would like to shed some light on multi-class subgroup discovery approach and its usefulness in rule learning framework with respect to the context of propositional logic as well as first-order logic.

Poster Abstract P06:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

A Comparison of CNF with CRF in Named Entity Recognition task

Kei Uchiumi KUCHIUMI@YAHOO-CORP.JP
Keigo Machinaga KMACHINA@YAHOO-CORP.JP
Toshiyuki Maezawa TMAEZAWA@YAHOO-CORP.JP
Toshinori Satou TOSHSATO@YAHOO-CORP.JP
R&D Unit, Yahoo Japan Corporation

Abstract

We compared Conditional Neural Fields (CNFs) (Peng et al., 2009) with Conditional Random Fields (CRFs) in a Japanese Named Entity Recognition (NER) task. CRFs are widely used for sequence labeling in natural language processing. CRFs use a linear potential function to represent the relationship between input features and an output label. On the other hand, CNFs use a non-linear potential function by adding one middle layer between input and output layers.

Regarding English NER tasks, it is known that Semi-Markov CRFs (Semi-CRFs) outperform the conventional CRFs. Semi-CRFs are a method of learning segmentation. Thus we added Semi-Markov CNFs (Semi-CNFs) to the comparison. We utilized CRF++ 0.54 and cnf-0.2.3 (http://code.google.com/p/cnf/) as implementations of CRFs, CNFs and Semi-CNFs. Semi-CNFs have never been used and compared in NER tasks, to our knowledge. It has been reported that Semi-CRF performance improves by adding features used in conventional CRFs (Andrew, 2006). So in our experiment, we added them to Semi-CNFs.

For evaluation, we made a corpus from CRL NE data and Kyoto University Text Corpus 3.0. CRL NE data was prepared for IREX (Information Retrieval and Extraction Exercise) and it has about 19,000 NEs in 1,174 articles of the Mainichi newspaper(1995). Kyoto University Text Corpus 3.0 is a tagged corpus of the Mainichi newspaper(1995). There are many common sentences between CRL NE data and Kyoto University Text Corpus. We added CRL's NE labels in IOB2 format to Kyoto University Text Corpus. In CRL NE data, The label 'OPTIONAL' is used for NEs that are difficult to annotate mannually. So We didn't use them for evaluation. We used 5-fold cross validation. For evaluation, we used F-measure, i.e. the weighted hermonic mean of precision and recall. Our experiment showed Semi-CNFs (87.28), CNFs (86.45) and CRFs (86.01) in F-measure. These results revealed that Semi-CNFs are the highest of the 3 methods.

Keywords: sequence labeling , conditional random fields, conditional neural fields

References

G. Andrew. A hybrid markov/semi-markov conditional random field for sequence segmentation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 465–472. Association for Computational Linguistics, 2006.

Jian Peng, Liefeng Bo, and Jinbo Xu. Conditional neural fields. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1419–1427. 2009.

Poster Abstract P07:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

Multiscale-Bagging with Applications to Classification

Masayoshi Aoki(Tokyo Institute of Technology) Takafumi Kanamori (Nagoya University) Hidetoshi Shimodaira (Tokyo Institute of Technology)

We consider the problem of prediction from a learning dataset $L = \{(x_t, y_t), t = 1, \ldots, n\}$ where the y's are the class labels. The present work is intended to find the true class label as the sample size $n \to \infty$. We attempt to find the true class label by the bagging method of Breiman (1996).

Ordinary bagging takes bootstrap samples $L^{*1}, \ldots, L^{*B}$ for prediction where B is the number of bootstrap resamplies of Efron (1979). After the resampling, we make a predictor for each bootstrap sample and decide the class label by a majority of these predictors. In this bagging procedure, we can obtain the frequency of the output of each class. The frequency is called the bootstrap probability (BP).

In this study, we use this BP as the confidence level of the null hypothesis. That is, we think of BP as an approximation of the p-value of null hypothesis which the true class label is y for input x. If BP falls below the significance level α, we reject the hypothesis of the class label and set the unrejected classes as candidates of the true class.

However, it is known that the BP has a bias. So, we propose the multiscale-bagging which uses the multiscale bootstrap algorithm of Shimodaira (2004, 2008). The multiscale bootstrap corrects the bias of the BP, and calculates the approximately unbiased p-values (AU). We use this AU in place of BP.

We performed a simulation study for evaluating the BP and AU. Unbiased p-values should satisfy that,

$$P(p < \alpha) = \alpha \qquad 0 < \alpha < 1,$$

that is, the distribution of unbiased p-values is uniform on $(0, 1)$ when x is on the true decision boundary. We check whether the BP and AU satisfy this condition.

References

Breiman, L. Bagging Predictors, *Machine Learning*, 24, 123-140, 1996

Efron, B. Bootstrap methods: Another look at the jackknife, *Annals of Statistics*, 7, 1-26, 1979

Shimodaira, H. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling, *Annals of Statistics*, 32, 2616-2641, 2004

Shimodaira, H. Testing Regions with Nonsmooth Boundaries via Multiscale Bootstrap, *Journal of Statistical Planning and Inference*, 138, 1227-1241, 2008

Poster Abstract P08:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

Contrasting Correlations by an Efficient Double-Clique Search Method

Aixiang Li and Makoto Haraguchi
Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan.
E-mail: aixiang,makoto@kb.ist.hokudai.ac.jp

Data mining has been applied to a broad range of activities that attempt to discover new information from existing databases. Contrast set mining is one of the most well-studied problems when given two or more comparable databases; additionally, correlation mining has also been paid much more attention in this decade. In this poster, we consider a problem of contrasting correlations over two databases (DBs). We try to find the itemsets consisting of items that are correlated at a low level in the one DB, but correlated at a medium level in the other DB.

Naturally, our correlation covers both positive and negative ones on the basis of considering the absence and presence of items in POS transactions (where the items are sales items) or in documents (where the items are words). Thus, each item can be considered as a categorical variable taking Boolean values. In general, correlation coefficient is useful for identifying linear functional dependence between random variables, but is poor at handling the case of categorical variables. The chi-squared value can be a candidate measure for the correlatedness of categorical variables. However we need to compare correlations of an itemset over two databases with different sizes. Because chi-squared value tends to be infinity as the sample size increase, it is not meaningful to compare chi-squared values over two different databases. For this reason, we introduce an extended mutual information (called k-way mutual information) to measure the degree of the correlation among items in an itemset so that the difference of database sizes does not make an influence on the contrasted values.

For two items, A and B, their correlation are needless to say calculated by the standard mutual information $I(A;B)$. However, as is well known, given the condition, the third item C, the 3-way extended mutual information, $I(A;B;C)$, shows higher value, even when there exists no dependence between A and B. Therefore, we consider the correlation not only between two items but also among three or more items extendedly. And then the correlation of k itemset, $X = \{x_1, \ldots, x_k\}$, is measured by the k-way mutual information, $I(X) = I(x_1; \ldots; x_k)$. Based on the k-way mutual information measure, our task is to find the itemsets whose correlation at one DB is low and the correlation at the other DB is medium. The problem can be defined as: Given items set $I = \{i_1, i_2, ..., i_n\}$, two databases DB_1 and DB_2, correlation constraints (upper bounds) δ_1 for DB_1 and δ_2 for DB_2 such that $\delta_1 < \delta_2$, correlation increase ratio $r\%$, find subsets $X \subseteq I$, subject to $I_{DB_1}(X) < \delta_1$, $I_{DB_2}(X) < \delta_2$ and $(I_{DB_2}(X) - I_{DB_1}(X))/I_{DB_1}(X) \geq r\%$.

To find the itemsets, the standard method is to enumerate all possible itemsets, computing their mutual informations and then comparing them. As the number of itemsets in the itemset lattice is very huge, the cost of the standard method is high, we have to develop some pruning rules to reject useless ones. It is well known that the k-way mutual information increases monotonically as the set X grows to larger sets. More precisely, if $I(A;B) > \delta$, then $I(A;B;C) > \delta$ holds. Based on the property, we introduce a new graph theoretic technique to reduce the candidate sets that violate the correlation constraints. In fact, we construct two graphs: one graph G_1 for DB_1, whose edges are drawn when the correlation between a pair of items is less than δ_1, and the other G_2 for DB_2 when the correlation between a pair of items is less than δ_2. By using the two graphs with the different types of edges, we search double cliques (cliques in both G_1 and G_2), and then compute their correlations in two databases. The itemsets (double cliques) satisfying the both correlation constraints and having a significant change rate of correlations are extracted.

Of course, the rare (or frequent) items or itemsets in extracted results are not useful. To avoid them, we consider a refined support constraint in searching itemsets. In this poster, an item is considered as a variable with boolean values. This leads to a fact that there are 2^k supports that corresponding to 2^k events for a set of k items. When the number of rare events is large, it is often the case that some included items are redundant and the combination of k items is not meaningful. Therefore, on the 2^k supports, we apply a refined form of support constraint that is downward closed to prune some useless sets efficiently.

Based on these strategies, we developed an algorithm based on depth-first double-clique search. We demonstrate its effectiveness by testing it on Nikkei POS data and BankSearch Web document dataset. It is showed that some interesting combinations of sales items or words are extracted, and the contrasting correlation problem can be solved efficiently.

Poster Abstract P09:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

Model-induced Regularization

Shinichi Nakajima NAKAJIMA.S@NIKON.CO.JP
Nikon Corporation, Tokyo 140-8601, Japan

Masashi Sugiyama SUGI@CS.TITECH.AC.JP
Tokyo Institute of Technology and JST PRESTO, Tokyo 152-8552, Japan

Abstract

When the *Bayesian* estimation is applied to modern probabilistic models, an *unintentional* strong regularization is often observed. We explain the mechanism of this effect, and introduce relevant works.

Suppose we are given i.i.d. samples $\{x_1, \ldots, x_n \in \mathbb{R}\}$ taken from Gaussian model with the mean parameter $u \in \mathbb{R}$:

$$p(x) = \mathcal{N}(x; u, 1^2). \quad (1)$$

Assuming Gaussian prior, $p_u(u) = \mathcal{N}(u; 0, c_u^2)$, where $c_u^2 > 0$ is a variance hyperparameter, we can perform Bayesian estimation, controlling c_u^2 as a regularization constant. What if we set c_u^2 to a large value ($c_u^2 \to \infty$)? The answer may be trivial; we get an unregularized estimator. (More accurately, the mode of the Bayesian predictive distribution coincides to the maximum likelihood (ML) estimator.)

Suppose next the following model:

$$p(x) = \mathcal{N}(x; ab, 1^2). \quad (2)$$

Here, the parameters are $a, b \in \mathbb{R}$, whose product corresponds to the parameter u in the original model (1). Let us assume Gaussian priors on a and b: $p_a(a) = \mathcal{N}(a; 0, c_a^2), p_b(b) = \mathcal{N}(b; 0, c_b^2)$. Will we similarly get an unregularized estimator of $u = ab$ when $c_a^2, c_b^2 \to \infty$?

The answer is NO. The estimator tends to be strongly regularized. We call this effect model-induced regularization (MIR), since it is inherent in the model likelihood function.

Actually, Eq.(2) is a special case of the matrix factorization model, and therefore, MIR explains the empirically observed superiority (Salakhutdinov & Mnih, 2008) of full-Bayesian estimation over maximum a posteriori (MAP) estimation. Here, note that MIR is caused by *density non-uniformity* of distribution functions in the parameter space, and therefore observed only when at least one parameter is integrated out. (No parameter is integrated out in MAP.) Other popular models in machine learning, e.g., mixture models and hidden Markov models, also have a similar structure to Eq.(2), which induces MIR.

The origin of MIR can be explained in terms of the Jeffreys prior (Jeffreys, 1946), with which the two models, (1) and (2), give the equivalent estimation. Another explanation has been done in the context of visual recognition (Freeman, 1994). Although the idea of the Jeffreys prior is widely known, people seem to underestimate the strength of this effect.

In our poster, we will explain why MIR occurs. Then, works that relate MIR with *singularities* of probabilistic models are introduced. A powerful procedure for quantitative evaluation of MIR has been developed, and applied to various models (Watanabe, 2009). Theoretical analysis has been extended to the variational Bayesian (VB) approximation. We will also introduce works that clarified the strength of MIR when VB is applied (Nakajima & Sugiyama, 2010).

References

Freeman, W. (1994). The Generic Viewpoint Assumption in a Framework for Visual Perception. *Nature*, *368*, 542–545.

Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London. Series A.* (pp. 453–461).

Nakajima, S., & Sugiyama, M. (2010). Implicit Regularization in Variational Bayesian Matrix Factorization. *ICML2010.*

Salakhutdinov, R., & Mnih, A. (2008). Bayesian Probabilistic Matrix Factorization using Markov Chain Monte Carlo. *ICML 2008.*

Watanabe, S. (2009). *Algebraic geometry and statistical learning.* Cambridge, UK: Cambridge University Press.

Poster Abstract P10:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

Slice Sampling on Chinese Restaurant Process

Takaki Makino
Division of Project Coordination, University of Tokyo
5-1-5 Kashiwa-no-ha, Kashiwa-shi, Chiba 277-8568 Japan
mak@scint.dpc.u-tokyo.ac.jp

We propose a method for applying slice sampling [1] to a predictive distribution obtained from Chinese Restaurant Process [2]. To perform correct slice sampling on the mixture of finite points of probability mass and probability distribution, our proposed method first decides the probability distribution to be sampled probabilistically, and perform slice sampling on the distribution. This sampling method can be combined with conditional simultaneous draw sampler [5] to construct an efficient sampler for more complex distributions, such as infinite HMM [6].

Slice sampling [1] is an Markov chain Monte Carlo sampling technique that draws samples from an unnormalized probability distribution. It is known that slice sampling on a discrete distribution, such as a multinomial distribution, effectively accelerates sampling by reducing the number of candidates considered in a sampling process [3]. Consider sampling from distribution $q(x) = \frac{f(x)p(x)}{Z}$, where $p(x)$ is a multinomial distribution over integer $\{1 \ldots N\}$, $f(x)$ is a non-negative function that is expensive to be evaluated, and Z is the normalization constant, i.e., $Z = \sum_x f(x)p(x)$. For efficient sampling, we want to avoid the normalization constant Z, which involves N times of evaluations of the function $f(x)$. In slice sampling, we first sample the auxiliary variable $u \sim \mathrm{Uniform}(0, p(x'))$ using the previous sample x', and draw the next sample from distribution $\tilde{q}(x) = \frac{f(x)I(p(x)>u)}{\tilde{Z}}$, where I is the indicator function, whose value is 1 if the condition is true and 0 otherwise, and $\tilde{Z} = \sum_{x:p(x)>u} f(x)$. When $p(x')$ is high (which is likely to occur), the condition $p(x) > u$ is likely to be false for many candidates, and the number of required evaluations of $f(x)$ in one sampling step is reduced. This is particularly effective when $p(x)$ is subject to a Dirichlet process, because we can avoid possibly infinite number of evaluations.

Theoretically, we can apply the Slice sampling technique to a predictive distribution of Chinese Restaurant Process. Consider an unknown distribution $G \sim \mathrm{CRP}(\alpha, H)$, where α is the hyperparameter and H is a base measure, and we observed n samples from G, which consists of m discrete values, $x_1, \ldots, x_m$, and the value x_i appears n_i times $(1 \le i \le m, \sum n_i = n)$. Then the predictive distribution of G is given as the following: $P = \sum_{i=1}^{m} \frac{n_i}{n+\alpha}\delta_{x_i} + \frac{\alpha}{n+\alpha}G_0$, where δ_x is a point distribution function on x.

However, in many cases, the support of H consist of an infinite number of possible values, each of which has infinitesimal probability mass compared to δ_x. If we directly apply slice sampling to P and the previous sample x' is equal to observed sample x_i, we will obtain a sample from one of such values only when the auxiliary variable u is infinitesimally small, that is expected to be probability zero.

If H does not have probability mass on any observed value x_i, we can solve this problem by gathering the infinite number of infinitesimal probabilities into a special valu 0, and perform slice sampling on $\widetilde{P} = \sum_{i=1}^{m} \frac{n_i}{n+\alpha}\delta_{x_i} + \frac{\alpha}{n+\alpha}\delta_0$; if the previous sample x' does not match to any of observed samples x_i, we consider as if the previous sample was 0, and if the obtained sample is 0, we replace the result x with a random sample from the base measure H. However, if H have non-zero probability for any observed value x_i, this sampling scheme is not correct. This happens in particular when we want to apply slice sampling on hierarchical Chinese restaurant process [4].

We propose a new method that performs sampling to obtain the probability distribution to be sliced. We first decide x_0, which is the candidate value corresponding to the value from H, as followings: if the previous sample x' does not match to any of the observed values x_i, let x_0 be x', and otherwise, obtain x_0 by a random sampling from the base measure H. After that, we perform slice distribution $\widehat{P} = \sum_{i=1}^{m} \frac{n_i}{n+\alpha}\delta_{x_i} + \frac{\alpha}{n+\alpha}\delta_{x_0}$; this is easily implemented by treating the two cases, when $x_0 = x_i$ for some i and when x_0 does not match to any x_i. We confirmed that correctness of this sampling method by showing that the method satisfies detailed balance.

References

[1] Radford Neal. Slice sampling. *Annals of Statistics*, 31:705–741, 2003.
[2] D. Aldous. Exchangeability and related topics. In *École d' été de Probabilités de Saint-Flour XIII–1983, Lecture notes in mathematics 1117*, pages 1–198. Springer Verlag, 1985.
[3] J. van Gael, Y. Saatci, Y. Teh, and Z. Ghahramani. Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pages 1088–1095, 2008.
[4] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
[5] Takaki Makino, Shunsuke Takei, Daichi Mochihashi, Issei Sato, and Toshihisa Takagi. Conditional simultaneous draws from hierarchical Chinese restaurant processes. In *Proceedings of Nonparametric BayesWorkshop 2009*, 2009.
[6] Matthew J. Beal, Zoubin Ghahramani, and Carl E. Rasmussen. The infinite hidden Markov model. In *Advances in Neural Information Processing Systems*, volume 14, pages 577–584. MIT Press, 2002.

Poster Abstract P11:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

Interactive Behavior Adaptation Through Dialogue Based on Bayesian Network

Saifuddin Md Tareeq*
The Graduate University for Advanced Studies
Department of Informatics, National Institute of Informatics
2-1-2 Hitotsubashi, Choyoda Ku, Tokyo
smtareeq@nii.ac.jp

Tetsunari Inamura
The Graduate University for Advanced Studies and
National Institute of Informatics
2-1-2 Hitotsubashi, Choyoda Ku, Tokyo
inamura@nii.ac.jp

***Abstract*— To advance robotics and incorporate robots into our daily lives, natural and intuitive approaches must be developed that allow new skills to be taught in a timely manner. Teaching and learning is an approach especially suited for robotic application with non-technical users. In this paradigm the robot is shown examples of the desired behavior and infers the demonstrator's latent control policy. To learn the behavior strategy, conventional methods observes sets of sensor input and command output, extracts meaningful relation between the sensor and commands using statistical methods. But the performance of the learning strongly depends on the quality of the dataset of sensor and command. When the dataset included significant data, the learning would be a success; however it is difficult to obtain significant data for human-robot interaction in real world, because the robots basically store the dataset in every process cycle. For example, when a user kept operating same command in same situation, the statistical learning procedure tends to output the frequent command even though the sensor is not the frequent but rare. To select the rare command for rare observation, the system should ignore insignificant frequent data to avoid bad learning quality. We proposed a technique to manage experience data with evaluation of significance of the dataset based on a concept of change in degree of confidence. For a small change in the degree of confidence, the situation is considered familiar and thus data is regarded as an insignificant for learning, so that data will be discarded. On the other hand when the change in the degree of confidence is larger the situation is considered unfamiliar and thus data is regarded as significant for learning and used for learning. In addition to the learning of policy and detecting and adapting to the changes in policies, in this paper we present that interactive learning has been integrated with our system so that the robot can make interaction with the user by asking question or clarifying situation when it has low confidence. We adopted Bayesian networks to represent the policy, because it can incorporate prior knowledge and causal interaction of sensor and command can be represented even though the observation of the user command is not well conducted and also it can output a degree of belief for behavior decision based on observation of sensor as evidence. Conventional simple belief calculation based on frequency of the dataset causes the problem that the system tends to output the most frequent command even though sensor input for rare situation is given, when the dataset observed continuously during the human-robot interaction. The problem arises because the prior probability is calculated using the numbers of observations. This factor also causes another problem that the robot cannot adapt rapidly to changeable policies of the user. We think the concept of checking change in the degree of confidence is also effective for this problem. We adopted Dirichlet distribution to evaluate the significance of data. The Dirichlet distribution represents not only event probability among several propositions, but also degree of confidence for the output probability just referring a set of number of observation for the propositions. The system calculates the degree of confidence before and after the current observation. The change in the two degrees of confidence can be regarded as the importance of the observation to the learning process. We developed a teaching and learning system using Bayesian network that incorporated our concept. In an example experiment, a user operates a mobile robot using a joystick controller. The robot observes the given motor command and distance sensor information obtained by eight sonar sensor mounted on the front side. In our previous paper we investigated the feasibility of the proposed method by experiments in which the robot could learn policy and also rapidly adapt to the changes in the user policy. The result showed the proposed method could ignore insignificant dataset to avoid obtaining inappropriate policy cause from too much frequent dataset. In this paper we conduct and discuss an experiment in which the robot interact with the user by asking questions or clarifying situation when it have low confidence.**

Poster Abstract P12:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

Maximum Volume Clustering

Gang Niu[1,3] Bo Dai[2] Lin Shang[1] Masashi Sugiyama[3]
[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]NLPR/LIAMA, Institute of Automation, Chinese Academy of Science
[3]Department of Computer Science, Tokyo Institute of Technology

Introduction

To the best of our knowledge, *Maximum Margin Clustering* (MMC)—which maximizes the margin between two opposite clusters—is the first clustering algorithm directly connecting to statistical learning theory. For this reason, MMC has been extensively investigated recently. However, the *large margin principle* (LMP) is not the only way to go. There is also a *large volume principle* (LVP). Roughly speaking, machine learning algorithms based on LVP should prefer hypotheses in some large-volume equivalence classes.

In this work, we propose a novel model for clustering called *Maximum Volume Clustering* (MVC), which serves as a prototype partitioning the data into two clusters based on LVP. Given the samples X_n, we construct an $n \times n$ symmetric positive definite matrix Q that contains pairwise information about X_n, and then an Q-dependent hypothesis space $\mathcal{H}_Q$. If there is a measure on $\mathcal{H}_Q$, namely the *power*, then we can talk about the *likelihood* or *confidence* of each equivalence class. Similarly to the *margin* in MMC, the notion of *volume* can also be regarded as an estimation of the power. Therefore, the larger the volume is, the more confident we are of the data partition. Thus we consider the partition lying in the equivalence class with the maximum volume as the best partition.

Similarly to other clustering approaches, the optimization problem involved in our MVC is NP-hard, so we introduce two approximation schemes: a soft-label MVC algorithm based on *sequential quadratic programming* and a hard-label one based on *semi-definite programming*. We show that these two approximations can be reduced to spectral clustering and MMC in special cases. Hence the proposed MVC model may be regarded as a natural extension of existing spectral and large margin approaches. We also establish finite sample stability and an error bound for the soft-label MVC method. Experiments show that the proposed MVC approach is promising.

Algorithm

The primal problem of the *Soft-Label Maximum Volume Clustering* is

$$\min_{\mathbf{h}\in\mathbb{R}^n} -2\|\mathbf{h}\|_1 + \gamma \mathbf{h}^\top Q \mathbf{h} \quad \text{s.t. } \|\mathbf{h}\|_2 = 1.$$

At the t-th iteration, the subproblem at the current solution $(\mathbf{h}_t, \eta_t)$ is

$$\min_{\mathbf{p}_t\in\mathbb{R}^n} \mathbf{p}_t^\top(\gamma Q - \eta_t I)\mathbf{p}_t + 2(\gamma Q \mathbf{h}_t - \operatorname{sgn}(\mathbf{h}_t))^\top \mathbf{p}_t$$
$$\text{s.t. } 2\mathbf{h}_t^\top \mathbf{p}_t + \mathbf{h}_t^\top \mathbf{h}_t = 1,\ |\mathbf{p}_t^\top \mathbf{1} + \mathbf{h}_t^\top \mathbf{1}| \le b. \quad (1)$$

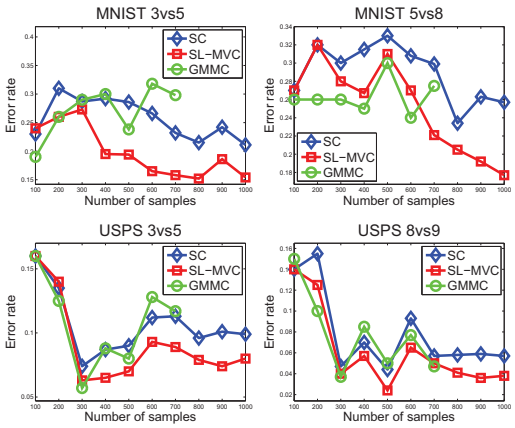
In our experiments we use an initialization $\mathbf{h}_0 = \operatorname{sgn}(\mathbf{v}_2 - \frac{1}{n}\mathbf{v}_2^\top \mathbf{1}\mathbf{1})/\|\operatorname{sgn}(\mathbf{v}_2 - \frac{1}{n}\mathbf{v}_2^\top \mathbf{1}\mathbf{1})\|_2$ and $\eta_0 = -0.001$, where $\mathbf{v}_2$ is the second smallest eigenvector of Q.

Algorithm 1: SL-MVC (SQP version)

input : stop criterion ϵ, matrix Q, regularization parameter γ, class balance parameter b
output: soft response vector $\mathbf{h}_{t+1}$

initialize $\mathbf{h}_0$ and η_0, and $t := -1$;
repeat
  $t := t + 1$; optimize (1) to obtain $\mathbf{p}_t$;
  update $\mathbf{h}_{t+1} := \mathbf{h}_t + \mathbf{p}_t$;
  update $\eta_{t+1} = (\gamma Q \mathbf{h}_{t+1} - \eta_t \mathbf{p}_t - \operatorname{sgn}(\mathbf{h}_t))^\top \mathbf{h}_t / (\mathbf{h}_t^\top \mathbf{h}_t)$;
until $\|\mathbf{h}_{t+1} - \mathbf{h}_t\|_2^2 + \|\eta_{t+1} - \eta_t\|_2^2 \le \epsilon$;

Experiments (partial)



Poster Abstract P13:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

Using Conditional Random Fields to validate observations in a 4W1H paradigm

Leon F. Palafox LEON@HLAB.IIS.U-TOKYO.AC.JP
Electric Engineering Department
University of Tokyo

Laszlo A. Jeni LASZLO@HLAB.IIS.U-TOKYO.AC.JP
Electric Engineering Department
University of Tokyo

Hideki Hashimoto HASHIMOTO@IIS.U-TOKYO.AC.JP
Electric Engineering Department
University of Tokyo

Abstract

Intelligent spaces final goal has been, since its inception, to help users to perform their daily tasks in the most effective way. To do this, a good system able to detect and classify human activity is needed. In the iSpace, we perform this classification using a large number of sensors attached both to the humans and the objects in the environment. We classify the sensors output using the 4W1H paradigm in which each of them provides specific information to a set of 5 variables (Where, What, Who, When and How) that can describe the current situation of the space.

To do the 4W1H classification we used specialized sensors, capable of providing information regarding each variable; systems equipped with linear accelerometers, gyroscopes and magnetometers provide an extensive reading of the movements for the How and What variables, for the Who, we used RFID tags attached to the users, and for the When and Where, we used the localized information (IP and Time) from the computer that was, in the reading time, closer to the user.

Yet, these systems are prone to failure, since the sensing must be exhaustive and relies on very sensitive accelerometers attached to the elements in the environment, the sensors in the objects may provide a misfiring if there is a brief movement that involves objects not related to the current activity. For example, if while typing a keyboard, accidentally a cup is moved, the system may classify the current activity as the user having a sudden sip of water and react accordingly to it by pouring some more water.

We propose a method to test and prevent these misfirings using conditional random fields (CFR), since they offer a training algorithm for sequential data that allows us to train the likelihood of the states in the system given that we know each state related observations. In our specific case, the observations of our system will be the 4W1H provided from the sensors (which are not random, thus the use of CFR against Markov Random Fields)and the states will be latent random variables that define the situation in the environment. Using the CFR we will know whether a current sequence of states is likely to occur in the system and if an unlikely sequence of observations happens, we can easily detect and classify it as a misfiring.

Keywords: Conditional Random Field, 4W1H, Human Activity Recognition

Poster Abstract P14:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

Multiscale Bagging with Applications to Classi cation and Active Learning

Hidetoshi Shimodaira[1] (shimo@is.titech.ac.jp)
Takafumi Kanamori[2] (kanamori@is.nagoya-u.ac.jp)
Masayoshi Aoki[1] (aoki6@is.titech.ac.jp)
Kouta Mine[1] (mine7@is.titech.ac.jp)

[1] Department of Mathematical and Computing Sciences, Tokyo Institute of Technology
[2] Department of Computer Science and Mathematical Informatics, Nagoya University

Abstract

We propose multiscale bagging as a modi cation of the bagging procedure. In ordinary bagging of Breiman (1996), the bootstrap resampling of Efron (1979) is used for generating bootstrap samples. We replace it with the multiscale bootstrap algorithm of Shimodaira (2002, 2004, 2008). In multiscale bagging, the sample size m of bootstrap samples may be altered from the sample size n of learning dataset. This is called the m out of n bootstrap in statistical literature, but our multiscale bootstrap is very different in the choice of m.

For assessing the output of a classi er, we compute bootstrap probability of class label; the frequency of observing a speci ed class label in the outputs of classi ers learned from bootstrap samples. A scaling-law of bootstrap probability with respect to a scale parameter

$$\sigma^2 = \frac{n}{m} \tag{1}$$

has been developed in connection with the geometrical theory of Efron and Tibshirani (1998).

We consider two different ways for using multiscale bagging of classi ers. The rst usage is to construct a con dence set of class labels, instead of a single label. The second usage is to nd inputs close to decision boundaries in the context of query by bagging (Abe and Mamitsuka, 1998) for active learning. The statistical theory proves that an appropriate choice of m is $m = -n$, i.e., $\sigma^2 = -1$, for the rst usage, and $m = \infty$, i.e., $\sigma^2 = 0$, for the second usage. For implementing this idea, we de ne a normalized bootstrap probability by

$$p_{\sigma^2}(\mathbf{x}, y) = \Phi\Big(\sigma\Phi^{-1}(\ \ _{\sigma^2}(\mathbf{x}; y))\Big), \tag{2}$$

where $\ \ _{\sigma^2}(\mathbf{x}, y)$ is the bootstrap probability for input $\mathbf{x}$ and output y, and $\Phi(\cdot)$ is the CDF of $N(0, 1)$. We then compute $p_{\sigma^2}(\mathbf{x}, y)$ for several $m > 0$ values, say, $m = n/2, n, 2n$, and extrapolate it to either

$$\sigma^2 = -1, \quad \text{or} \quad \sigma^2 = 0. \tag{3}$$

References

- B. Efron and R. Tibshirani. The problem of regions. *Annals of Statistics*, 26: 1687–1718, 1998.
- H. Shimodaira. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Annals of Statistics*, 32:2616–2641, 2004.
- H. Shimodaira. Testing regions with nonsmooth boundaries via multiscale bootstrap. *Journal of Statistical Planning and Inference*, 138:1227–1241, 2008

Poster Abstract P15:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

Adjustment for multiple hypotheses testing in comparative classification studies

Daniel Berrar BERRAR.D.AA@M.TITECH.AC.JP
Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology
4259 Nagatsuta, Midori-ku
Yokohama 226-8502, Japan

Abstract

Comparative classification studies that include more than two models entail the problem of multiple hypotheses testing. Here, we investigated this problem for fully specified classifiers that are compared on a small-sample, independent test set. Our analysis included only classifiers with a known true performance and therefore with known pair-wise differences. We evaluated the significance of these differences using McNemar's test and confidence intervals for performance measures, with and without corrections for multiple testing. We compared five classifiers using a real-world data set and two Monte Carlo experiments, simulating 10000 comparative studies each. In our experiments, we observed that the confidence intervals based on Quesenberry and Hurst with adjustments based on Holm's method provided for an acceptable Type I error while the Type II error was not overly elevated. The Holm-adjusted intervals based on Quesenberry and Hurst are recommended for comparing multiple fully specified classifiers on an independent test set, specifically when this test set is relatively small, as in the present study. Comparative studies generally include more than two models, and the resulting multiplicity problem needs to be adequately addressed to avoid erroneous conclusions about the differences in performance.

Keywords: classification, multiple testing, McNemar's test, Quesenberry and Hurst interval

Poster Abstract P16:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

Inference in Latent Conditional Models: the Computational Complexity Analysis and a Comparative Study of Solutions

Xu Sun XUSUN@MIST.I.U-TOKYO.AC.JP
Department of Mathematical Informatics
University of Tokyo

Hisashi Kashima KASHIMA@MIST.I.U-TOKYO.AC.JP
Department of Mathematical Informatics
University of Tokyo

Takuya Matsuzaki MATUZAKI@IS.S.U-TOKYO.AC.JP
Department of Computer Science
University of Tokyo

Abstract

Latent conditional models have become popular recently in both natural language processing and vision processing communities. Since efficient inference on traditional models, like CRFs, are well-solved by using dynamic programming, it seems to many people that the inference in latent conditional models can also be efficiently solved by using similar techniques. To make the situation clear for the upcoming studies on this direction, in this paper we analyzed the computational complexity of the inference problem in latent conditional models, and show that this problem is an NP-hard problem. To the extent of our knowledge, this is the first proof of the NP-hardness of the latent conditional models, even in a simple linear chain case. It also indicates that more complicated structured latent conditional models (e.g., tree-structured latent models for syntactic parsing) is also NP-hard. Besides the analysis on the computational complexity, we also made a comparative study on the various inference methods based on approximation techniques or heuristics. Our experiments demonstrate that the bounded version of the latent-dynamic inference outperforms other alternative methods, and with quite fast inference speed in practice.

Poster Abstract P17:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

Improving Graph-based Semi-Supervised Learning by Feature Space Transformation

Yu-Shi Lin[1,2] D96008@CSIE.NTU.EDU.TW
[1]*Institute of Information Science*
Academia Sinica
Taipei, Taiwan
[2]*Dept. of Computer Sci. and Info. Eng.*
National Taiwan University
Taipei, Taiwan

Chun-Nan Hsu[1,3] CHUNNAN@ISI.EDU
[3]*Information Sciences Institute*
Univ. of Southern California
Marina del Rey, CA, USA

Abstract

Semi-supervised learning (SSL) is important when labeled training examples are rare or expensive. Graph-based SSL is one of the most promising SSL methods but its performance depends heavily on the graph structure constructed on top of the feature space. Here we present a feature space transformation method based on the spectral graph theory to automatically improve the performance of graph-based SSL. We used the graph transition energy as the objective function so that in effect, minimizing this objective will optimally pull together data points with the same labels while push apart those with different labels. Thus, our method can transform the feature space by re-weighting edges of the graph so that its regularized graph transition energy can be minimized. We argue that minimizing this new objective function will lead to a better graph structure that in turns implies better SSL performance about classifying unlabeled data. We derive a lower bound for this new objective function. The quality of the transformation can be estimated by comparing the minimization with a theoretical lower bound.

To evaluate the effectiveness of our feature space transformation method, we combine our method along with several graph-based SSL methods including combinations of graph construction methods, b-nearest neighbors and b-matching, and label inference algorithms including k-nearest neighbor classifier and the Gaussian Random Field method. We investigated its performance for synthetic datasets and datasets of fluorescent microscopic cell images. SSL is particularly suitable for this application because of recent advances in high-throughput image-based assays which allow for rapid acquisition of a large number of cell images for analysis. However, labeled training data is rare and expensive to get and thus limited in quantity for supervised learning methods.We show that our method constantly improves the classification performance substantially no matter what combination is used. Moreover, we show that improvement can be accomplished even though the initial graph is imbalanced and irregularly constructed.

Keywords: Semi-supervised learning, spectral graph theory, feature space transformation

Poster Abstract P18:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

Image Annotation via Multi-Instance Learning with Pyramid Graph Kernel

Zhi Nie, Guiguang Ding, and Chunping Li

Software school, Tsinghua University, Beijing, 100084, China
E-mail: niez07@mails.tsinghua.edu.cn

Abstract— In this work, we invented the Multi-Layer Node Kernel to discover the probable positive patterns in a bag and make these instances play the key role in determining its label. For Image annotation, the label of a bag may be determined by more than one instance. In response to this problem, we designed the edge kernel to capture the pairwise co-occurrence relationship essential to the target concept. Finally, the Multi-layer Node Kernel and edge kernel were optimally combined to form the Pyramid Graph kernel through multiple kernel learning methods. The experiments were carried on Corel Image dataset.

Keywords— image annotation, multi-instance learning, kernel methods

1. Construction of Pyramid Graphs

The basic idea to solve the problem of capturing the presumably complex distribution of positive instances and co-occurrence relationship essential to the target concept is to build a pyramid of graphs. Each graph in the pyramid is formed by clustering training instances, with each node representing a cluster. 2^l nodes are formed at l^{th} level of the pyramid. The weight assigned to each node is dependent on how many instances it contains from positive training bags. So the weight of the node reflects its probability of being a positive pattern. For unseen bags, the instances are assigned to the cluster as follows:

$$\Phi_{g^l}(x_{ij}) = \arg\min_{k=1,\cdots,2^l} \left\| v_k^l - x_{ij} \right\|_2$$

in which the v_k^l is the cluster center of the k^{th} node at the l^{th} level of the pyramid, x_{ij} is the j^{th} instance in the bag X_i. There is an edge connecting two nodes or the node itself if two different instances in a bag are mapped to them or itself. The weight of an edge, however, is determined by its appearance in both positive and negative bags since though it is possible that positive patterns can appear in negative bags, the cases that co-occurrence of positive patterns in negative bags are rather rare. The graph construction process can be illustrated by figure 1.

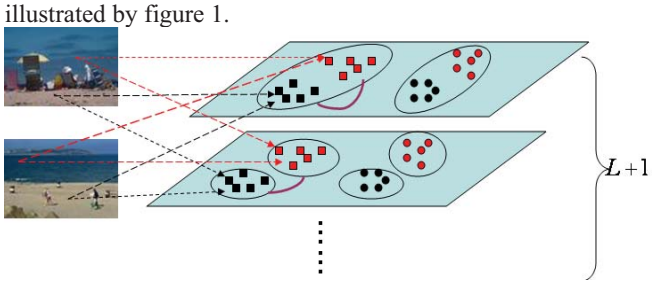


Figure 1: illustration of construction of pyramid graphs. For clarity, the 0th level of the pyramid is not shown

2. Multi-layer Node Kernel

Let X_i and X_j be two multi-instance bags, G^l be the graph at the l^{th} level of the pyramid and L+1 be total levels of the pyramid .The Multi-Layer Node Kernel is defined as

$$k^l_{node}(X^i, X^j) = \frac{\sum_{a=1}^{n_i}\sum_{b=1}^{n_j} w^l_{\Phi_{g^l}(x_{ia})} w^l_{\Phi_{g^l}(x_{jb})} k(x_{ia}, x_{jb})}{\sum_{a=1}^{n_i}(w^l_{\Phi_{g^l}(x_{ia})} + \sigma)\sum_{b=1}^{n_j}(w^l_{\Phi_{g^l}(x_{jb})} + \sigma)};$$

$$K_{node}(X^i, X^j) = \sum_{l=0}^{L} \frac{1}{2^{L-l}} k^l_{node}(X^i, X^j);$$

where $k(\cdot,\cdot)$ is defined as the Gaussian RBF kernel. σ is a small value to keep the denominator from being zero. The illustration for the case of L=2 is shown in figure 2.

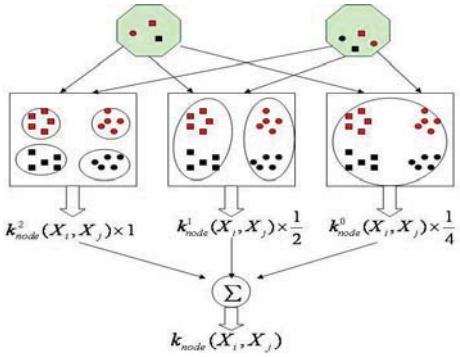


Figure 2: illustration of the Multi-Layer Node Kernel when L = 2.

3. Pyramid Graph Kernel

The key to Pyramid Graph Kernel lies in defining a kernel for pairs of edges. When comparing two edges, we are concerned with whether the patterns two edges connect are same. Also, since at lower levels of pyramid, the instances in all the training bags are split into many clusters, one cluster at the upper levels of the pyramid graphs may consists of several clusters at the lower levels. Thus, we define the distance between two edges as the Earth Mover's Distance between the two pair of nodes. The edge kernel is defined much like the multi-layer node kernel. The positive semi-definiteness of Multi-Layer Node Kernel and edge kernel can be formally proved. Finally, the Pyramid Graph Kernel can be derived by combining the two kernels as follows $K(X_i, X_j) = \mu_1 K_{node}(X_i, X_j) + \mu_2 K_{edge}(X_i, X_j)$ s.t. $\mu_1, \mu_2 \geq 0, \mu_1 + \mu_2 = 1$;

4. Experiments

The experiments were conducted on Corel Image 2000 data set and repeated 5 times. The mean results is shown as follows

Table 1. Mean Classification Accuracy

Kernel type	L	Mean accuracy
Multi-Layer Node Kernel	9	70.4
Pyramid Graph Kernel	9	71.9

As to the selection of L, we recommend that some criterions from information theory, like the minimum message length criterion, be utilized.

Poster Abstract P19:
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010.

Proximity in Large Bipartite Graphs with Unsupervised Auxiliary Information

Rudy Raymond, Yuta Tsuboi
IBM Research – Tokyo

Hisashi Kashima, Issei Sato
The University of Tokyo

1. SUMMARY

Many interactions in the real world can be expressed as bipartite graphs whose nodes can be partitioned into two parts (called *left* and *right* nodes) such that all edges of the graph link nodes from different parts. There are many natural examples of such graphs. For example, an author-conference bipartite graph whose nodes represent scientists or conferences, and whose edges link nodes corresponding to scientists to those of conferences, thus representing the contributed-to relationship. Moreover, a general graph can be turned into a bipartite one by copying all of its nodes into left and right nodes, and adding corresponding links between left and right nodes appropriately. For this reason, we can also consider a host-host bipartite graph whose nodes represent Web hosts, and whose edges represent hyperlinks of Web pages on the hosts. Proximity scores of nodes on such graphs have many important applications in recommendation, ranking, link prediction, etc.

In many such typical bipartite graphs although the total number of nodes is large (for example, the number of scientific paper authors registered in DBLP from 1990 to 2008 is more than 490,000), they are often *skewed*, so the number of nodes in one part is relatively small compared to the other (there are only about 4,000 conferences registered in DBLP for the same period). Skewed bipartite graphs have special properties that can be exploited for designing efficient algorithms to compute proximity scores from their topological structures. One of them is the so-called Random Walk with Restart (or, RWR for short)(Tong et al., KAIS 2008), that calculates a proximity score of node v to node u from the steady-state probability of reaching v from u by a random walk. The principle of RWR is similar to the well-known random-surfer model of PageRank (Page et al., Stanford 1998) on general graphs, however, the scores of RWR on bipartite graphs are easier to compute, especially, when the number of left and right nodes is highly unbalanced. This has sparked widespread interest on measuring proximities with RWR, even for dynamic bipartite graphs (Tong et al., SDM 2008). However, only little is known about how to take into account information other than the link structure for proximity measurements.

Quite recently, a fast algorithm for proximities that incorporates supervised auxiliary information for general graphs was proposed in (Tong et al., ICDM 2008), where the supervised auxiliary information was regarded as binary information and used to refine the link structure of the underlying graph. This was done with techniques taking into account the user's favourable preference by adding new links between the corresponding user's node and its marked positive nodes, and the unfavourable preference, by adding new links between the marked negative nodes and their neighbors with a special node without outlinks (a sink node). The techniques require careful selection of parameters for the weight of links between positive and negative nodes as well as for the selection of neighboring nodes.

At the same time, one can also obtain other types of information that take continous values representing the degree of similarities between nodes in the graph. For example, the similarities of Web hosts in the host-host bipartite graphs can be calculated from the inner product of their keyword features, or, similarities of users can be measured from the overlaps of their tracks, friends or social tags, and so on. Those scores of similarities are obviously not discrete, and thus, present us with a challenge as to how to incorporate them for better proximity scores.

In this paper, we propose a new approach for refining the proximity scores of RWR with such unsupervised auxiliary information. Our approach is based on the *graph label propagation* for deriving a minimization problem that guides the RWR to assign similar scores to similar nodes (and diverging scores to dissimilar nodes), without explicitly changing the structure of the underlying graph. The auxiliary information only gives the (dis)similarity scores of partial nodes in the graph and does not give their preferred order explicitly, and hence the term unsupervised. We designed our approach so that it still retains the advantages of RWR on bipartite graphs. Its computational complexity is at most the same as that of the original RWR. Therefore, we believe that our approach will be useful for enriching the applicability and the effectiveness of the RWR. For this reason, we also include some interesting experimental results on applying proximity scores, from both the RWR and our new approach, in labeling Web spam hosts and in link prediction for large bipartite graphs using real-world Web-spam host graphs and social network datasets.

To summarize, our contributions in this paper are threefolds: (1) We present a novel approach of using unsupervised auxiliary information to adjust the proximity scores of RWR. (2) We describe an efficient procedure to obtain the adjusted scores incorporating the auxiliary information from the original scores of RWR. (3) We present experimental results using proximity scores of the RWR and the adjusted scores for labeling Web spam hosts and for predicting links in large bipartite graphs. We confirmed that the auxiliary information is helpful in refining the effectiveness of proximity scores of RWR.