ICML2011 Unsupervised and Transfer Learning Workshop

Daniel L. Silver DANNY.SILVER@ACADIAU.CA
Acadia University, Canada

Isabelle Guyon ISABELLE@CLOPINET.COM
Clopinet, California, USA

Graham Taylor GWTAYLOR@CS.NYU.EDU
New York University, USA

Gideon Dror GIDEON@MTA.AC.IL
Academic College of Tel-Aviv-Yaffo, Israel

Vincent Lemaire VINCENT.LEMAIRE@ORANGE-FTGROUP.COM
Orange Labs, France

Editor: Neil Lawrence

Abstract

We organized a data mining challenge in "unsupervised and transfer learning" (the UTL challenge) followed by a workshop of the same name at the ICML 2011 conference in Bellevue, Washington[1]. This introduction presents the highlights of the outstanding contributions that were made, which are regrouped in this issue of JMLR W&CP. Novel methodologies emerged to capitalize on large volumes of unlabeled data from tasks related (but different) from a target task, including a method to learn data kernels (similarity measures) and new deep architectures for feature learning.

Keywords: transfer learning, unsupervised learning, metric learning, kernel learning, unlabeled data, challenges

1. Introduction

Unsupervised learning considers the problem of discovering regularities or structure in unlabeled data (*e.g.,* finding sub-manifolds or clustering examples) based on a representation of the domain. Transfer learning considers the use of prior knowledge (such as labeled training examples, or shared features) from one or more source tasks when developing a hypothesis for a new target task. While human beings are adept at transfer learning using mixtures of labeled and unlabeled examples, even across widely disparate domains, we have only begun to develop machine learning systems that exhibit the combined use of unsupervised learning and knowledge transfer.

To foster greater research in this area we organized a international challenge on Unsupervised and Transfer Learning that culminated in a workshop of the same name at the ICML-2011 conference in Bellevue, Washington, on July 2, 2011. This workshop addressed a question of fundamental and practical interest in machine learning: the development and

[1] http://clopinet.com/isabelle/Projects/ICML2011/.

assessment of methods that can generate data representations (features) that can be reused across domains of tasks.

This edition of JMLR W&CP presents the challenge results and a collection of outstanding contributed articles on the subject of transfer learning and unsupervised learning. This paper and the edition focuses on unsupervised and transfer learning for classification problems based on real-valued feature representations that are related more closely to data mining tasks. Methods of transfer learning have also been investigated for reinforcement learning (Ramon et al., 2007; Taylor and Stone, 2007), however these are outside the scope of this edition.

2. Overview of Transfer Learning and Unsupervised Learning

2.1. Transfer Learning

Transfer learning refers to use of knowledge for one or more source tasks to develop efficiently a more accurate hypothesis for a new target task. Transfer learning has most frequently been applied to sets of labeled data that have a supervised target value for each example. For instance, there would be significant benefit in using an accurate diagnostic model of one disease to develop a diagnostic model for a second related disease for which you have few training examples. While all learning involves generalization across problem instances, transfer learning emphasizes the transfer of knowledge across domains, tasks, and distributions that are similar but not the same. Inductive transfer has gone by a variety of names: bias learning, learning to learn, machine life-long learning, knowledge transfer, transfer learning, meta-learning, and incremental, cumulative, and continual learning.

Research in inductive transfer began in the early 1980s with discussions on inductive bias, generalization and the necessity of heuristics for developing accurate hypotheses from small numbers of training examples (Mitchell, 1980; Utgoff, 1986). This early research suggested that the accumulation of prior knowledge for the purposes of selecting inductive bias is a useful characteristic for any learning system. Following the first major workshop on inductive transfer (NIPS1995 Workshop, 1995) a series of articles were published in special issues of Connection Science (Lorien Pratt (Editor), 1996) and Machine Learning (Pratt and Sebastian Thrun (Editors), 1997), and a book entitled "Learning to Learn" (Thrun and Lorien Y. Pratt (Editors), 1997) .

Since that time, research on inductive transfer has occurred using traditional machine learning methods (Caruana, 1997; Baxter, 1997; Silver and Mercer, 1996; Heskes, 2000; Thrun and Lorien Y. Pratt (Editors), 1997; Bakker and Heskes, 2003; Ben-David and Schuller, 2003), statistical regression methods (Greene, 2002; Zellner, 1962; Breiman and Friedman, 1998), Bayesian methods involving constraints such as hyper priors (Allenby and Rossi, 1999; Arora et al., 1998; Bakker and Heskes, 2003), and more recently kernel methods such as support vector machines (SVMs) (Jebara, 2004; Allenby and Rossi, 2005). All of these approaches rely upon the development of a hypothesis for a target task under a constraint or regularization that characterizes a similarity or *relatedness* to one or more source tasks. In 2005, a second major workshop on inductive transfer occurred at NIPS. Papers from this workshop can be found in (Silver and Bennett, 2008) as well as at (NIPS2005 Workshop, 2005).

More recently, there has been work on inductive transfer in the areas of self-taught learning (Raina et al., 2007b), transductive learning (Arnold et al., 2007), *context-sensitive* multiple task learning (Silver et al., 2008), the learning of model structure (Niculescu-Mizil and Caruana, 2007), unsupervised transfer learning [Yu,Wang], and a variety of methods that mix unsupervised and supervised learning to be discussed in greater detail below.

2.2. Unsupervised Learning

Unsupervised learning refers to the process of finding structure in unlabeled data resulting in new data representations (including feature representations) and/or clustering data into categories of similar examples, based on such representations (Hinton and Sejnowski, 1999). The unlabeled data distinguishes unsupervised learning from supervised learning and reinforcement learning. Important recent progress has been made in purely unsupervised learning (Smola et al., 2001; Bengio et al., 2003; Globerson and Tishby, 2003; Ghahramani, 2004; Lawrence, 2005; Luxburg, 2007). However, these advances tend to be ignored by practitioners who continue using a handful of popular algorithms like PCA and ICA (for feature extraction and dimensionality reduction), and K-means, and various hierarchical clustering methods for clustering (Jain et al., 1999).

2.3. Combining Unsupervised and Transfer Learning

It is often easier to obtain large quantities of unlabeled data from databases and sources on the web, for example images of unlabeled objects. For this reason the idea of using unsupervised learning in combination with supervised learning has attracted interest for some time. Semi-supervised learning is a machine learning approach that is halfway between supervised and unsupervised learning. In addition to the labeled data for a given task of interest, the algorithm is provided with unlabeled data for the *same* task - typically a small amount of labeled data and a large amount of unlabeled data (Blum and Mitchell, 1998). Note that these approaches usually assume that the categories of the unlabeled data, even though unknown to the learning machine, are the same as the categories of the labeled data, *i.e.,* that the "tasks" are the same.

In contrast, in the transfer learning setting, the unlabeled data does not need to come from the same task. There has been considerable progress in the past decade in developing cross-task transfer using both discriminative and generative approaches in a wide variety of settings (Pan and Yang, 2010). These approaches include multi-layer structured learning machines from the "Deep Learning" family such as convolutional neural networks, Deep Belief Networks, and Deep Boltzmann Machines (Bengio, 2009; Gutstein, 2010; Erhan et al., 2010), sparse coding (Lee et al., 2007; Raina et al., 2007a), and metric or kernel learning methods (Bromley et al., 1994; WU et al., 2009; Kulis, 2010). The "Learning to learn" and "Lifelong Learning" veins of research have continued to provide interesting results in both machine learning and cognitive science in terms of short-term learning with transfer and long-term retention of learned knowledge (Silver et al., 2008). These references include recent evidence of the value of combining unsupervised generative learning with transfer learning to generate a rich set of representation (features) upon which to build related supervised discriminative tasks. The goal of the challenge we organized was to perform an evaluation of unsupervised and transfer learning algorithms free of inventor bias to help to identify and popularize algorithms that have advanced the state of the art.

3. Overview of the UTL Challenge

Part of the ICML workshop was devoted to the presentation of the results of the Unsupervised and Transfer Learning challenge (UTL challenge Guyon et al., 2011a,b). The challenge, which started in December 2010 and ended in April 2011, was organized in 2 phases. The aim of **Phase 1** was to benchmark **unsupervised learning** algorithms used as preprocessors for supervised learning, in the context of transfer learning problems. The aim of **phase 2** was to encourage researchers to exploit the possibilities offered by new cutting-edge cross-task transfer learning algorithms, which **transfer supervised learning knowledge from task to task**.

To that end, the competitors were presented with five datasets illustrating classification problems from different domains: handwriting recognition, video processing, text processing, object recognition, and ecology. Each dataset was split into 3 subsets: development, validation, and final evaluation sets. In phase 1, all subsets were provided without labels to the participants. The labels remained known only to the organizers throughout the challenge. The goal of the participants was to produce the best possible data representation for the final evaluation data. This representation was then evaluated by the organizers on supervised learning classification tasks by training and testing a linear classifier on subsets of the final evaluation data, such than a learning curve would be produced. The evaluation metric was the area under the learning curve, which is a means of aggregating performance results over a range of number of training examples considered.

To avoid the possibility of participants selecting their model based on final evaluation set performance, the final results remained secret until the end of the challenge. Rather, feedback was provided on-line during the challenge on the performance obtained on validation data, and the final evaluation set data was used only for the final ranking. For both phases, the participants could either submit a data representation (for validation data and final evaluation data) or a matrix of similarity between examples (a kernel). Hence, the competition was equivalently a data representation learning challenge and a kernel learning challenge.

In contrast with a classical evaluation of unsupervised learning as a preprocessing, the three subsets (development, validation, and final evaluation sets) were **not drawn from the same distribution**. In fact, they all had different sets of class labels. Picture for instance a problem of optical character recognition (OCR), the development set could contain only lowercase alphabetical letters, the validation set could contain uppercase letters, and the final evaluation set, digits and symbols. This setting is typical of real world problems in which there is an abundance of data available for training from a source domain, which is distinct from the target domain of interest. For instance, in face recognition, there is an abundance of pictures from unknown strangers that are available on the Internet, compared to the few images of your close family members that you care to classify. The development set represents a source domain whereas the validation and final evaluation sets represent alternative target domains on which different sets of tasks can be defined[2].

In the second phase of the challenge, a few labels of the development set were provided, offering to the participants the possibility of using supervised learning in some way to produce better data representations for the validation and final evaluation sets. The setting remained otherwise unchanged.

One of the main findings of this challenge is the power of unsupervised learning as a preprocessing tool. For all the datasets of the challenge, unsupervised learning produced results significantly better than the baseline methods (raw data or simple normalizations). The participants exploited effectively the feed-back received on the validation set to select the best data representations. The skepticism around the effectiveness of unsupervised learning is justified when no performance on a supervised task is available. However, unsupervised learning can be the object of model selection using a supervised task, similarly to preprocessing, feature selection, and hyperparameter selection. An interesting new outcome of this challenge is that the supervised tasks used for model selection can be distinct from the tasks used for the final evaluation. So, even though the learning algorithms are unsupervised, transfer learning is happening at the model selection level. This setting is related to the "self-taught learning" setting proposed in (Raina et al., 2007a). Another interesting finding is that, perhaps the development set is not useful at all. The winners of phase 1 did not use it. They devised a method to select a cascade of preprocessing steps to be used to produce a new kernel. The same cascade was then applied to produce the kernel of the final evaluation set(Aiolli, 2012). The importance of the degree of resemblance of the validation task and final task remains to be determined.

In phase 1, there was a danger of overfitting by trying too many methods and relying too heavily on the performance on the validation set. One team for instance overfitted in phase 1, ranking 1st on the validation set, but only 4th on the final evaluation set. Possibly, criteria involving both the reconstruction error and the classification accuracy on the validation tasks may be more effective for model selection. This should be the object of further research. In phase 2, the participants had available "transfer labels" for a subset of the development data (for classification tasks distinct from the classification tasks of the validation set and the final evaluation set). Therefore, they had the opportunity to use such labels to devise transfer learning strategies. The most effective strategy seems to have been to use the transfer labels for model selection again. None of the participants used those labels for learning.

Overall, an array of algorithms were used (Aiolli, 2012; Le Borgne, 2011; Liu et al., 2012; Mesnil et al, 2012; Saeed, 2011; Xu et al, 2011), including linear methods like Principal Component Analysis (PCA), and non-linear methods like clustering (K-means and hierarchical clustering being the most popular), Kernel-PCA (KPCA), non-linear auto-encoders and restricted Bolzmann machines (RBMs). A general methodology seems to have emerged. Most top ranking participants used simple normalizations (like variable standardization and/or data sphering using PCA) as a first step, followed by one or several layers of non-linear pro-

[2] In this paper, we call "domain" the input space (*e.g.,* a feature vector space) and we call "task" the output space (represented by labels for classification problems). We use the adjective "source" for an auxiliary problem, for which we have an abundance of data (*e.g.,* pictures of strangers in the Internet), and "target" for the problem of interest (*e.g.,* pictures of family members).

cessing (stacks of auto-encoders, RBMs, KPCA, and/or clustering). Finally, "transduction" played a key role in winning first place: either the whole preprocessing chain was applied directly to the final evaluation data (this is the strategy of Fabio Aiolli who won first place in phase 1, Aiolli, 2012); or alternatively, the final evaluation data, preprocessed with a preprocessor trained on development+validation data, was post-processed with PCA (so-called "transductive PCA" used by the LISA team, who won the second phase, Mesnil et al, 2012).

4. Overview of Proceedings

The following provides an overview of the workshop proceedings including the tutorials, invited presentations, challenge winner articles and other refereed articles submitted to the workshop.

4.1. Tutorials

The workshop provided two foundational tutorials included in this proceeding. The morning tutorial covered *Deep Learning of Representations for Unsupervised and Transfer Learning* with Yoshua Bengio from the Université de Montréal (Bengio, 2012). Deep learning algorithms seek to exploit the unknown structure in the input distribution in order to discover good representations, often at multiple levels, with higher-level learned features defined in terms of lower-level features. The paper focusses on why unsupervised pre-training of representations using autoencoders and Restricted Boltzmann Machines can be useful, and how it can be exploited in the transfer learning scenario, where we care about predictions on examples that are not from the same distribution as the training distribution.

The afternoon tutorial entitled *Towards Heterogeneous Transfer Learning* was presented by Qiang Yang, Hong Kong University of Science, co-author of an authoritative review of transfer learning (Pan and Yang, 2010). Transfer learning has focused on knowledge transfer between domains with the same or similar input spaces. The heterogeneous transfer approach considers the ability to use knowledge from very different task domains and input spaces. The authors demonstrated heterogeneous transfer learning between text classification and image classification domains even when there are no explicit feature mappings provided. They explained that the key is to identify and maximize the commonalities among the internal structures (features) of the different domains.

4.2. Challenge Winner Articles

Three teams were presented awards at the workshop for their winning performances on the UTL Challenge and their authorship. This section will summarize the papers describing these winning entries.

The first place award for phase 1 of the UTL challenge (unsupervised learning) as well as the Pascal2 best challenge paper award for phase 1 went to Fabio Aiolli and his paper *Transfer Learning by Kernel Meta-Learning* (Aiolli, 2012). Recently, there have been a number of researchers who have investigated the problem of finding a good kernel matrix for a task. This is known as kernel learning. Kernel learning can be transformed into a semi-supervised learning problem by using a large set of unlabeled data and a smaller set of labeled data. The paper presents a novel approach to transfer learning based on kernel learning with both labeled and unlabeled data. Starting from a basic kernel, the method attempts to learn chains of kernel transforms that produce good kernel matrices for a set of source tasks. The same sequence of transformations are then applied to learn the kernel matrix for a target task. The application of this method to the five datasets of the Unsupervised and Transfer Learning (UTL) challenge produced the best results for the first phase of the competition.

The LISA team of the Université de Montréal, Canada, ranked first in the second phase of the UTL challenge (transfer learning), and their paper entitled *Unsupervised and Transfer Learning Challenge: A Deep Learning Approach* (Mesnil et al, 2012) won the Pascal2 best challenge paper award for phase 2. The LISA team demonstrated the usefulness of Deep Learning architectures to extract internal representations from a large set of unlabeled training examples. This is accomplished by introducing gradually network layers trained in an unsupervised way using the feature representation of lower layers. The final representation is then used to train a simple linear classifier with a small number of labeled training examples.

The team "1055a" of Chuanren Liu, Jianjun Xie, Hui Xiong, and Yong Ge of CoreLogic and Rutgers University won the second place award for phase 1 of the UTL challenge (unsupervised learning) and came in at third place in phase 2 (transfer learning) (Liu et al., 2012). Their paper entitled *Stochastic Unsupervised Learning on Unlabeled Data* was also selected for inclusion in these proceedings. The paper introduces a stochastic unsupervised learning method that performs as a preprocessing K-means clustering on principal components extracted from the raw unlabeled data. This removes the effect of noise and less-relevant features improving the methods robustness. The approach utilizes a stochastic process to combine multiple clustering assignments on each data point to alleviate over-fitting.

We also include in the supplemental material poster presentations and technical reports of work, which was not yet ready for publication, but shows interesting new directions of research:

The team of Zhixiang Xu (Airbus), Washington University in St. Louis, who took third place in phase 1 of the UTL challenge presented a poster entitled "Rapid Feature Learning with Stacked Linear Denoisers" (Xu et al, 2011). They investigated unsupervised pre-training of deep architectures as feature generators for shallow classifiers. They implemented a computationally efficient algorithm that mimics stacked denoising auto-encoders (SdAs). Their feature transformation improves the results of SVM classification, sometimes outperforming SdAs and deep neural networks.

Mehreen Saeed (Aliphlaila team), fourth place phase 2 UTL challenge), FAST, Pakistan, communicated a technical report entitled "Use of Representations in High Dimensional Spaces for Unsupervised and Transfer Learning Challenge" (Saeed, 2011). The author shows how manifold learning and simple similarity kernels can be used to get good results.

Yann-Aël Le Borgne (Tryan team, fourth place in second ranking of phase 2 UTL challenge), VUB, Belgium, showed a poster entitled "Supervised Dimensionality Reduction in the Unsupervised and Transfer Learning 2011 Competition" (Le Borgne, 2011). The author presented preliminary results on a technique making use of all three subsets provided for each dataset (development, validation, and final evaluation datasets) to assign labels to samples. The author then uses partial least square (PLS) to extract features of interest.

4.3. Fundamentals and Algorithms

The UTL workshop papers gathered in these proceedings follow two main axes. One axis ranges **from theory to application** of transfer learning and the other **from supervised learning, to unsupervised learning** and hybrid approaches. The following summarizes each of articles along those two dimensions.

In *Autoencoders, Unsupervised Learning, and Deep Architectures*, Pierre Baldi of UC Irvine, investigates the theoretical underbelly of autoencoders (Baldi, 2012). He presents a mathematical framework for the study of both linear and non-linear autoencoders - particularly the non-linear case of a Boolean autoencoder. He shows that learning with a Boolean autoencoder is equivalent to a clustering problem that can be solved in polynomial time when the number of clusters is small and becomes NP complete when the number of clusters is large. The framework sheds light on the connections between different kinds of autoencoders, their learning complexity and their composition in deep architectures. The paper brings together much of the theory on autoencoders, clustering, Hebbian learning, and information theory.

Joachim Buhmann et al. of ETH, Zurich, present a paper entitled, "Information Theoretic Model Selection for Pattern Analysis" (Buhmann et al, 2012). The authors propose a method of model and model-order selection for unsupervised data clustering based on information theory. Their approach ranks competing pattern cost functions according to their ability to extract context sensitive information from noisy data with respect to the hypothesis class. Sets of approximative solutions serve as a basis for an information theoretic communication protocol. Inferred models maximize the so-called "approximation capacity" that measures the mutual information between training data patterns and test data patterns, each of which have been made optimally "coarse" through the controlled addition of random noise. The approach is demonstrated using a Gaussian mixture model.

4.4. Supervised, Unsupervised and Transductive Approaches

4.4.1. Supervised

The workshop provided new insights into supervised learning approaches to transfer. Ruslan Salakhutdinov et al. of MIT, USA, presented their paper on *One-Shot Learning with a Hierarchical Nonparametric Bayesian Model* (Salakhutdinov et al., 2012). One-shot learning is the ability to develop a general classification model from a single training example. The authors develop a hierarchical Bayesian model that can transfer acquired knowledge from previously learned categories to a novel category, in the form of a prior over category means and variances. The model discovers how to group categories into meaningful super-categories and infer to which super-category a novel example belongs, and thereby estimate not only the new category's mean but also an appropriate similarity metric. The method is tested using the MNIST and MSR Cambridge image datasets and shown to perform significantly better than simpler hierarchical Bayesian approaches, discovering new categories in a completely unsupervised fashion.

In *Inductive Transfer for Bayesian Network Structure Learning*, Alexandru Niculescu-Mizil (NEC Laboratories America) and Rich Caruana (Microsoft Research) consider the problem of jointly learning the structures of Bayesian network models from multiple related datasets (Niculescu-Mizil and Caruana, 2012). They present an algorithm that simultaneously learns a multi-task Bayesian network structure by transferring useful information between the different datasets. The algorithm extends the heuristic search techniques used in traditional structure learning to the multi-task case by defining a scoring function for sets of structures (one structure for each dataset) and an efficient procedure for searching for a set of structures that has a high score across all tasks. The approach assumes that the true dependency structures of related problems are similar: the presence or absence of arcs in some of the structures provides evidence for the presence or absence of those same arcs in the other structures.

Kohei Hayashi and Takashi Takenouchi of the Nara Institute of Science and Technology, Japan, in their paper *Self-measuring Similarity for Multi-task Gaussian Process* extend work by Bonilla et al. (2008) on a multi-task Gaussian process framework (Hayashi et al, 2012). Their approach incorporates similarities between tasks based on the observed responses, which allows for the representation of much more complex data structures. The proposed approach is able to construct covariance matrices via kernel functions even when additional information such as example target values are available. The authors propose an efficient conjugate-gradient-based algorithm that implements the approach. The method is shown to perform the best to date on the Movielens 100k dataset.

4.4.2. Unsupervised

The workshop also provided a number of new approaches to transfer using unsupervised learning or combinations of supervised and unsupervised transfer learning. In *Clustering: Science or Art?*, Ulrike von Luxburg et al. examine whether the quality of different clustering algorithms can be compared by a general, scientifically sound procedure, which is independent of particular clustering algorithms (von Luxburg et al., 2012). They conclude that clustering should not be treated as an application-independent mathematical problem, but should always be studied in the context of its end-use. Different reasons for clustering bring with it different metrics for success. They argue that research spent on developing a "taxonomy of clustering problems" will be more fruitful than efforts spent on developing a domain independent clustering algorithm.

Preferably, high dimensional data, such as pixels of an image, are better described in terms of a small number of meta-features. In their paper *Unsupervised dimensionality reduction via gradient-based matrix factorization with two learning rates and their automatic updates* , Vladimir Nikulin and Tian-Hsiang Huang, of University of Queensland, Australia prescribe three related methods that combine to reduce noise while still capturing the essential features of the original data (Nikulin and Huang, 2012). The resulting features can then be used to do supervised classification. The proposed methods are demonstrated on the classification of gene expression data from cancer research where the number of labeled samples is relatively small compared to the number of genes in each sample.

4.4.3. Transductive

Ayan Acharya et al. report in their paper *Transfer Learning with Cluster Ensembles* a method of transferring learned knowledge from a set of source tasks when the target task has no labeled examples (Acharya et al, 2012). They present an optimization framework that applies the outputs of a cluster ensemble on a target task to moderate posterior probability estimates provided by classifiers previously induced on a related domain of tasks, so that the posterior probabilities are better adapted to the new context. This framework is general in that it admits a wide range of loss functions and classification/clustering methods. Empirical results on both text and hyperspectral data indicate that the proposed method can yield substantially superior classification results as compared to competing transductive learning techniques (Transductive SVM, Locally Weighted Ensemble).

4.5. Case studies

The final five papers from the workshop can be considered applications or case studies of unsupervised and transfer learning. The first paper, entitled *Transfer Learning in Computational Biology* applies multiple task learning to several problems in computational biology where the generation of training labels is often very costly (Widmer and Rätsch, 2012). The authors, Christian Widmer and Gunnar Raetsch, of MPI in Germany, received the Pascal2 best paper award at the workshop for this work. The paper presents two problems from sequence biology and uses regularization (SVM) based transfer learning methods, with a special focus on the case of a hierarchical relationship between tasks. The authors propose strategies to learn or refine a measure of task relatedness so as to optimize the transfer from source to target task.

In *Transfer Learning in Sequential Decision Problems: A Hierarchical Bayesian Approach*, Aaron Wilson et al, of Oregon State University, show that transfer is doubly beneficial in reinforcement learning where the agent not only needs to generalize from sparse experience, but also needs to discover good opportunities to learn in the first place (Wilson et al., 2012). They show that the hierarchical Bayesian framework can be readily adapted to sequential decision problems and provides a natural formalization of transfer learning.

Transfer Learning for Auto-gating of Flow Cytometry Data. by Gyemin Lee et al. of the University of Michigan, Ann Arbor, apply transfer learning to flow cytometry, a technique for rapidly quantifying physical and chemical properties of large numbers of cells (Lee et al., 2012). In clinical applications, flow cytometry data must be manually "gated"(scored) to identify cell populations of interest. The authors leverage existing datasets, previously gated by experts, to automatically gate a new flow cytometry dataset while accounting for biological variation. An empirical study demonstrates the approach by automatically gating lymphocytes from peripheral blood samples. The authors received the Pascal2 best student paper award for this work.

Philemon Brakel and Benjamin Schrauwen, of Ghent University, Belgium, use a hierarchical Bayesian logistic regression model to perform a binary document classification task in the paper *Transfer Learning for Document Classification: Sampling Informative Priors* Their approach estimates the covariance matrix of a multivariate Gaussian prior over the model parameters using a set of related tasks. Inference was done using a combination of Hybrid Monte Carlo and Gibbs sampling. They demonstrate that the obtained priors contain information that is beneficial for developing a model for document classification from small training sets.

Finally, in *Divide and Transfer: an Exploration of Segmented Transfer to Detect Wikipedia Vandalism*, Si-Chi Chin and W. Nick Street, of the University of Iowa, apply knowledge transfer methods to the problem of detecting Wikipedia vandalism (Chin and Street, 2012). Transfer is used to address the problem of small amounts of labeled data by leveraging unlabeled data and previously acquired knowledge from related source tasks. Avoiding negative transfer becomes a primary concern given the diverse nature of Wikipedia modifications that can occur. The proposed two segmented transfer approaches map unlabeled data from the target task to the most related cluster from the source task, classifying the unlabeled data using the most relevant learned models.

5. Summary

Challenges foster progress in particular scientific domains, but their specific formulation may bias research in too narrow ways. For that reason, our workshop invited diverse contributions on the theme of transfer learning, in addition to discussing the results of the challenge we organized. As a result, it is more difficult to draw general conclusions summarizing the enormous body of work that this represents. In some sense, transfer learning covers all the aspects of machine learning, with the only particularity that training data includes "source" domains and/or tasks that do are different from the "target" domains and/or tasks of interest. Within this general setting, many types of transfer learning formulations have been made. From our point of view, the most notable contribution of these proceedings is to demonstrate the effectiveness of recently proposed methods in the context of a wide variety of real world applications, both through the results of the challenge and other contributed papers. We hope that the mix of articles collected in this proceedings issue will spark further interest and curiosity in transfer learning. There is much work to be done in this area in terms of new computational learning theory and the application of existing algorithms and techniques.

Acknowledgements

We are very grateful to all authors for their submissions to the workshop and for their work in improving and polishing their articles for this proceedings issue. We would also like to thank the referees for their reviews and helpful comments to authors that have led to improvements in content and format. Finally, we are particularly grateful to the Editor-in-Chief, Neil Lawrence, for the opportunity to compile this proceedings issue and to the editorial staff of JMLR Workshop and Conference Proceedings for their kind assistance. Our generous sponsors and data donors and our dedicated advisors, beta testers listed on our challenge website (http://clopinet.com/ul) are gratefully acknowledged. This project is part of the DARPA Deep Learning program and is an activity of the Causality Workbench supported by the Pascal network of excellence funded by the European Commission and by the U.S. National Science Foundation under Grant N0. ECCS-0725746. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

A. Acharya et al. Transfer learning with cluster ensembles. In *ICML 2011 Unsupervised and Transfer Learning Workshop*. JMLR W&CP, this volume, 2012.

F. Aiolli. Transfer learning by kernel meta-learning. In *ICML 2011 Unsupervised and Transfer Learning Workshop*. JMLR W&CP, this volume, 2012.

G. M. Allenby and P. E. Rossi. Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89:57–78, 1999.

G. M. Allenby and P. E. Rossi. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

A. Arnold, R. Nallapati, and W. W. Cohen. A comparative study of methods for transductive transfer learning. In *In ICDM Workshop on Mining and Management of Biological Data*, 2007.

N. Arora, G. M. Allenby, and J. Ginter. A hierarchical bayes model of primary and secondary demand. *Marketing Science*, 17(1):29–44, 1998.

B. Bakker and T. Heskes. Task clustering and gating for bayesian multi-task learning. *Journal of Machine Learning Research*, 4:83–99, 2003.

P. Baldi. Autoencoders, unsupervised learning, and deep architectures. In *ICML 2011 Unsupervised and Transfer Learning Workshop*. JMLR W&CP, this volume, 2012.

J. Baxter. Theoretical models of learning to learn. *Learning to Learn*, pages 71–94, 1997.

S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. In *Proceedings of Computational Learning Theory (COLT)*, pages 185–192, 2003.

Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009. doi: 10.1561/2200000006. Also published as a book. Now Publishers, 2009.

Y. Bengio. Deep learning of representations for unsupervised and transfer learning. In *ICML 2011 Unsupervised and Transfer Learning Workshop*. JMLR W&CP, this volume, 2012.

Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, Nicolas Le Roux, and Marie Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In *NIPS03*, 2003.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT' 98 Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. Morgan Kaufmann Publishers, 1998.

L. Breiman and J. H. Friedman. Predicting multivariate responses in multiple linear regression. *Royal Statistical Society Series B*, 1998.

J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a ”siamese” time delay neural network. In *In NIPS Proc*, 1994.

J. Buhmann et al. Information theoretic model selection for pattern analysis. In *ICML 2011 Unsupervised and Transfer Learning Workshop*. JMLR W&CP, this volume, 2012.

R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

S.-C. Chin and W. Nick Street. Divide and transfer: an exploration of segmented transfer to detect wikipedia vandalism. In *ICML 2011 Unsupervised and Transfer Learning Workshop*. JMLR W&CP, this volume, 2012.

D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio. Why does unsupervised pre-training help deep learning? *JMLR*, 11:625–660, 2010.

Z. Ghahramani. *Unsupervised Learning*, volume 3176, pages 72–112. Springer-Verlag, Berlin, 2004.

A. Globerson and N. Tishby. Sufficient dimensionality reduction. *J. Mach. Learn. Res.*, 3:1307–1331, March 2003. ISSN 1532-4435. URL http://portal.acm.org/citation.cfm?id=944919.944975.

W. Greene. *Econometric Analysis, 5th Edition*. Prentice Hall, 2002.

S. M. Gutstein. *Transfer Learning Techniques for Deep Neural Nets*. PhD thesis, The University of Texas at El Paso, 2010.

I. Guyon, G. Dror, V. Lemaire, D. Silver, G. Taylor, and D. W. Aha. Analysis of the ijcnn 2011 utl challenge. *Neural Networks*, In press 2011a.

I. Guyon, G. Dror, V. Lemaire, G. Taylor, and D. W. Aha. Unsupervised and transfer learning challenge. In *The 2011 International Join Conference on Neural Networks*, pages 793–800, July 2011b.

K. Hayashi et al. Self-measuring similarity for multi-task gaussian process. In *ICML 2011 Unsupervised and Transfer Learning Workshop*. JMLR W&CP, this volume, 2012.

T. Heskes. Empirical bayes for learning to learn. In P. Langley, editor, *Proceedings of the International Conference on Machine Learning (ICML'00)*, pages 367–374, 2000.

G. E. Hinton and T. J. Sejnowski. *Unsupervised learning: foundations of neural computation*. Computational neuroscience. MIT Press, 1999. ISBN 9780262581684. URL http://books.google.ca/books?id=yj04Y0lje4cC.

A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review, 1999.

T. Jebara. Multi-task feature and kernel selection for svms. In *Proceedings of the International Conference on Machine Learning (ICML 04)*, pages 185–192, 2004.

B. Kulis. Icml tutorial on metric learning, 2010. URL http://www.cs.berkeley.edu/~kulis/icml2010_tutorial.htm.

N. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *JMLR*, 6:1783–1816, 2005.

Y.-A. Le Borgne. Supervised dimensionality reduction in the unsupervised and transfer learning 2011 competition, 2011.

G. Lee, L. Stoolman, and C. Scott. Transfer learning for auto-gating of flow cytometry data. In *ICML 2011 Unsupervised and Transfer Learning Workshop*. JMLR W&CP, this volume, 2012.

H. Lee, A. Battle, R.t Raina, and A. Y. Ng. Efficient sparse coding algorithms. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 801–808. MIT Press, Cambridge, MA, 2007.

C. Liu, J. Xie, Hui Xiong, and Y. Ge. Stochastic unsupervised learning on unlabeled data. In *ICML 2011 Unsupervised and Transfer Learning Workshop*. JMLR W&CP, this volume, 2012.

L. Pratt (Editor). Reuse of neural networks through transfer. *Connection Science*, 8(2), 1996.

U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, December 2007. ISSN 0960-3174. doi: 10.1007/s11222-007-9033-z. URL http://portal.acm.org/citation.cfm?id=1288822.1288832.

G. Mesnil et al. Unsupervised and transfer learning challenge: a deep learning approach. In *ICML 2011 Unsupervised and Transfer Learning Workshop*. JMLR W&CP, this volume, 2012.

T. M. Mitchell. The need for biases in learning generalizations. *Readings in Machine Learning*, pages 184–191, 1980. ed. Jude W. Shavlik and Thomas G. Dietterich.

A. Niculescu-Mizil and R. Caruana. Inductive transfer for bayesian network structure learning. *Journal of Machine Learning Research - Proceedings Track*, 2:339–346, 2007.

A. Niculescu-Mizil and R. Caruana. Inductive transfer for bayesian network structure learning. In *ICML 2011 Unsupervised and Transfer Learning Workshop*. JMLR W&CP, this volume, 2012.

V. Nikulin and T.-H. Huang. Unsupervised dimensionality reduction via gradient-based matrix factorization with two adaptive learning rates. In *ICML 2011 Unsupervised and Transfer Learning Workshop*. JMLR W&CP, this volume, 2012.

NIPS1995 Workshop. Learning to learn. http://plato.acadiau.ca/courses/comp/dsilver/NIPS95_LTL/transfer.workshop.1995.html, 1995.

NIPS2005 Workshop. Inductive transfer - 10 years later. http://iitrl.acadiau.ca/itws05/, 2005.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knoweledge and Data Engineering*, 22(10):1345–1359, October 2010.

L. Pratt and S. Thrun (Editors). Transfer in inductive systems. *Machine Learning*, 28(1), 1997.

R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: Transfer learning from unlabeled data. In *Proceedings of the Twenty-fourth International Conference on Machine Learning*, 2007a.

R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 759–766, New York, NY, USA, 2007b. ACM. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273592. URL http://doi.acm.org/10.1145/1273496.1273592.

J. Ramon, K. Driessens, and T. Croonenborghs. Transfer learning in reinforcement learning problems through partial policy recycling. *Proc. 18th European Conf. Machine Learning (ECML 2007)*, pages 699–707, 2007.

M. Saeed. Use of representations in high dimensional spaces for unsupervised and transfer learning challenge, 2011.

R. Salakhutdinov, J. Tenenbaum, and A. Torralba. One-shot learning with a hierarchical nonparametric bayesian model. In *ICML 2011 Unsupervised and Transfer Learning Workshop*. JMLR W&CP, this volume, 2012.

D. L. Silver and K. P. Bennett. Special issue on inductive transfer. *Machine Learning*, 73, 2008.

D. L. Silver and R. E. Mercer. The parallel transfer of task knowledge using dynamic learning rates based on a measure of relatedness. *Connection Science Special Issue: Transfer in Inductive Systems*, 8(2):277–294, 1996.

D. L. Silver, R. Poirier, and D. Currie. Inductive transfer with context-sensitive neural networks. *Machine Learning*, 73:323–336, 2008.

A. J. Smola, S. Mika, B. Schölkopf, and R. C. Williamson. Regularized principal manifolds. *JMLR*, 1:179–209, 2001.

M.E. Taylor and P. Stone. Cross-domain transfer for reinforcement learning. *Proc. 24th International Conf. Machine Learning (ICML 2007)*, pages 879–886, 2007.

S. Thrun and L. Y. Pratt (Editors). *Learning To Learn*. Kluwer Academc Publisher, Boston, MA, 1997.

P. E. Utgoff. *Machine Learning of Inductive Bias*. Kluwer Academc Publisher, Boston, MA, 1986.

U. von Luxburg, R. C. Williamson, and I. Guyon. Clustering: Science or art? In *ICML 2011 Unsupervised and Transfer Learning Workshop*. JMLR W&CP, this volume, 2012.

C. Widmer and G. Rätsch. Multitask learning in computational biology. In *ICML 2011 Unsupervised and Transfer Learning Workshop*. JMLR W&CP, this volume, 2012.

A. Wilson, A. Fern, and P. Tadepalli. Transfer learning in sequential decision problems: A hierarchical bayesian approach. In *ICML 2011 Unsupervised and Transfer Learning Workshop*. JMLR W&CP, this volume, 2012.

X.-M. WU, A. Man-Cho So, Z. Li, and S.-Y. R. Li. Fast graph laplacian regularized kernel learning via semidefinite quadratic linear programming. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1964–1972. Curran Associates, Inc., 2009.

Z. Xu et al. Rapid feature learning with stacked linear denoisers, 2011.

A. Zellner. An efficient method for estimating seemingly unrelated regression equations and tests for aggregation bias. *Journal of the American Statistical Association*, 57:348–368, 1962.