Learning with Multiple Views
Proposal for an ICML Workshop

Stefan Rüping[1], Tobias Scheffer[2]

[1] Dortmund University, LS Informatik 8, 44221 Dortmund, Germany
Phone +49 231 755-5104, Fax +49 231 755-5105
stefan.rueping@uni-dortmund.de
[2] Humboldt University, Department of Computer Science, Unter den Linden 6
10099 Berlin, Germany; Phone +4930 2093 3107, Fax +4930 2093 3045
scheffer@informatik.hu-berlin.de

1 Introduction

We propose to have a workshop on multi-view learning at the Twenty-Second International Conference on Machine Learning. Two main reasons lead us to the conclusion that the community would benefit from such a workshop.

- Multi-view learning is a natural, yet non-standard new problem setting; in brevity: learning from instances that have multiple independent representations. It is motivated by several interesting application areas; they include semi-supervised and unsupervised learning from linked objects, such as web pages and scientific literature.
- A recent result by Abney [1] suggests that there may be an underlying principle which gives rise to a family of new methods: The disagreement rate of two independent hypotheses upper-bounds the error rate of either hypothesis. By minimizing the disagreement rate on unlabeled data, the error rate can be minimized. In the last 2-3 years, several new methods have been proposed which appear to utilize this consensus maximization principle in one way or another. However, in many cases the contributors are not to the full extent aware of the relationships between their methods and a possible common underlying principle.

The workshop aims at bringing together researchers who are working on learning problems with multiply represented instances and consensus maximizing learning methods; our goals are to make the intrinsic structure of this field more clearly visible and to bring this interesting and rapidly developing area to the attention of additional researchers.

2 Background

Multi-view learning describes the setting of learning from data where observations are represented by multiple independent sets of features. A typical example is learning to classify web pages, where a web document can be described by either the words occurring on the page itself or the words contained in anchor text of links pointing to this page.

Multi-view learning methods have been studied by Yarowsky [37] and Blum and Mitchell [5]; they noticed that having multiple representations instead of combining all features into one view can improve classification performance when in addition to labeled examples, many unlabeled examples are available. The idea of the co-training approach is to train one learner on each view of the labeled examples and then to iteratively let each learner label the unlabeled examples it predicts with the highest confidence. Given independence between the learners, newly labeled examples from one learner may give the other learner new information to improve its model with. A theoretical foundation of multi-view learning has been given by Abney [1], who showed that given a certain type of independence between the learners, the disagreement of two learners gives an upper bound on the error rate. Unlabeled data can be used to minimize the disagreement between learners, and hence improves their combined accuracy. Many approaches have followed and extended the original co-training idea, e.g., [28, 18, 6, 27, 8]

But multi-view learning is not restricted to the semi-supervised case. Bickel and Scheffer [3] modified the co-training algorithm by replacing the class variable with a mixture coefficient to obtain a multi-view clustering algorithm. Multi-view clustering has also been investigated under the name of clustering multi-type [35] or multi-represented [23] objects. Gondek and Hofmann [20, 19] introduced the novel task of non-redundant clustering with the goal of obtaining a clustering that is non-redundant with respect to available background knowledge. This can be interpreted as a multi-view setting, where one view is explicitly defined by the available background knowledge. Several other unsupervised algorithms can also interpreted as multi-view approaches, e.g., [12, 37, 9, 13].

In a fully supervised setting, multi-view learning usually performs inferior to learning on the union of all views. However, certain non-standard learning tasks contain a multi-view component. Rüping [31] investigated the problem of learning interpretable models by augmenting an interpretable global model with independent, high-complexity local models. In this case, one way to ensure interpretability is to let the user restrict the set or number of features the global learner may use; i.e., to prefer a user-defined view. Tsochantiridis and Hofmann [33] investigated the task of polycategorical classification, i.e., learning with more than one label. Their algorithm utilizes dependencies among the labels, such that each predicted label can be interpreted as a specifically constructed view on the examples.

This discussion shows that multi-view learning has, explicitly or implicitly, been applied in many, very different approaches, although many authors do not seem to be aware of the multi-view aspect and the connections of their work to approaches from different fields. There are also several more general connections between multi-view learning and other learning tasks, such as transductive learning [21, 4, 22, 38, 8, 39], learning with the expectation maximization algorithm [11] and learning with multiple classifiers [34, 7, 36, 15].

3 Topic and Goals of the Workshop

The goal of this workshop will be to explore and structure the field of multi-view learning and develop a common framework for the previously unconnected approaches for learning with multiple views. Furthermore, the workshop will serve to more formally distinguish the field of multi-view learning from related fields of machine learning and show up similarities and differences. In particular, the following topics are relevant:

Analysis of algorithms: Many multi-view algorithms have been proposed in the literature. What are their similarities and differences, which are the important aspects for comparing such algorithms (e.g., assumptions about learners and data, learning tasks, spectrum from unsupervised over semi-supervised to supervised learning)? Can the different approaches be generalized and cast into a general framework? What is the best way to measures the confidence of a learner in co-training, the independence of hypotheses and the disagreement of learners?

Novel learning tasks: The utilization of multiple learners and hidden variables can be used to improve the learner on other performance criteria, for example interpretability [31], or adapt it to certain constraints [20] and learning tasks [12]. Which other criteria, constraints and learning tasks exist where multi-view learning is particularly well suited?

Practical aspects of the independence assumption: The core assumption of co-training is the independence of the base learners given the labels. There are several ways that can be used to ensure this independence, for example using multiple views, i.e., different descriptions of the data [5, 3, 6], using different hypothesis spaces in the base learners [31] or explicitly optimizing the independence in the algorithm [16, 20]. Are there other approaches? How do this approaches compare in practice? Are the assumptions of independence fulfilled in practice? What is the best way to measure the independence of learners on a certain data set? Can we tell whether multi-view learning will work before starting the learner?

Theoretical analysis: Multi-view algorithms have been theoretically analyzed with respect to PAC bounds [10], independence assumptions [1] and with respect to a formulation as optimizing an objective function [2]. However, theoretical analysis is far from finished. Open questions are: What happens if independence assumptions are only partially fulfilled? What can be said about convergence guarantees and speed? When can global optimal solutions be achieved? Are the algorithms statistically robust with respect to noise in the data?

Relation to other fields of machine learning: Related fields are learning with labeled and unlabeled data, e.g., transductive Support Vector Machines [21], graph mincuts [4], spectral graph partitioning [22] and learning within a regularization framework [38, 8] (see also [32] for an overview on semi-supervised learning). Another interesting relation exists to learning with multiple classifiers, such as voting, combination by order statistics [34], Meta-Level Learning [7], Stacking [36], and Boosting [15]. Learning with the expectation maximization algorithm [11] has also intensively been applied. What is the multi-view aspect of these approaches and what insights can multi-view learning and these fields gain from each other? Under which conditions does one approach perform better than the other, in particular in comparison with classical supervised learning?

Applications: Informative applications where multi-view approaches are particularly well suited for learning are also of interest for the workshop.

4 Intended Audience

The intended audience of the workshop are researchers in the field of machine learning and statistics with background in one of the following fields: multi-view learning, co-learning, clustering, transductive learning, learning theory, multi-classifier learning.

We plan on an audience of 20-30 participants. Many presentations will be invited but we will also issue an open call for papers. Potential participants are the PC members as well as the authors of the papers cited in this proposal.

5 Format and Publicity

We intend to organize a full-day workshop; the workshop will include (many) invited and (some) contributed presentations with a strong emphasis on discussions. The workshop proceedings will appear both online on the workshop website and as a technical report (we would be willing to print the proceedings unless the ICML organizers prefer to print workshop proceedings centralized); we will distribute the proceedings to the workshop participants free of charge.

6 Organizers, Program Committee

The workshop will be organized by Stefan Rüping and Tobias Scheffer.

Stefan Rüping is Research Associate at the AI unit at Dortmund University, working in the DFG Collaborative Research Center 475 on Reduction of Complexity for Multivariate Data Structures. His expertise with respect to the workshop is on EM learning interpretable models [31] and probabilistic classification [30, 29]. He has served as a PC member at ICDM 2004.

Tobias Scheffer is Assistant Professor at Humboldt-Universität zu Berlin, Germany. He is interested in statistical machine learning, learning from text, and applications to information retrieval and bioinformatics. Tobias Scheffer has been working on multi-view clustering and classification [3, 6, 25]. He has organized several workshops (including an ECML workshop on active learning and instance selection, and ECML workshops on text mining in bioinformatics), and has served as PC member and area chair at past and present ICML, ECML, and other conferences.

The following colleagues have already agreed to become a member of the program committee, should the workshop be accepted.

Steffen Bickel, Humboldt University [3].
Ulf Brefeld, Humboldt University [6].
Sanjoy Dasgupta, University of California, San Diego [10].
Johannes Fürnkranz, Darmstadt University [17].
Rayid Ghani, Accenture [27].
Thomas Hofmann, Brown University [33, 19, 20].
Thorsten Joachims, Cornell University [21, 22].
Kristian Kersting, Freiburg University [14].
Stan Matwin, University of Ottawa [24].
Ion Muslea, SRI [26].
Tom Mitchell, Carnegie Mellon University [5].
Bernhard Schölkopf, Max Planck Institute for Biological Cybernetics [38, 8].

References

1. Steven Abney. Bootstrapping. In *40th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference*, 2002.
2. Steven Abney. Understanding the yarowsky algorithm. *Computational Linguistics*, 30(3), 2004.
3. Steffen Bickel and Tobias Scheffer. Multi-view clustering. In *Proceedings of the IEEE International Conference on Data Mining*, 2004.
4. Avrim Blum and Shuchi Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 19–26, 2001.
5. Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Annual Conference on Computational Learning Theory (COLT-98)*, 1998.
6. Ulf Brefeld and Tobias Scheffer. Co-EM support vector learning. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.
7. Philip K. Chan and Salvatore Stolfo. Experiments in multistrategy learning by meta-learning. In *Proceedings of the second international conference on informationand knowledge management*, pages 314–323, Washington, DC, 1993.
8. O. Chapelle, B. Schölkopf, and J. Weston. Semi-supervised learning through principal directions estimation. In *ICML Workshop, The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, 2003.
9. Michael Collins and Yoram Singer. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.
10. Sanjoy Dasgupta, Michael L. Littman, and David McAllester. Pac generalization bounds for co-training. In *Advances in Neural Information Processing Systems (NIPS)*, 2001.

11. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Ser. B*, 39:1–38, 1977.
12. I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 89–98, 2003.
13. Gal Elidan and Nir Friedman. The information bottleneck em algorithm. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, 2003.
14. J. Fischer and K. Kersting. Scaled cgem: A fast accelerated em. In H. Blockeel N. Lavrac, D. Gamberger and L. Todorovski, editors, *Proceedings of the Fourteenth European Conference on Machine Learning (ECML-2003)*, pages 133–144, Cavtat, Croatia, Sep 2003.
15. Y Freund and R.E Schapire. Game theory, on-line prediction and boosting. In *Proceedings of the 9th Annual Conference on Computational Learnin g Theory*, pages 325–332, 1996.
16. Nir Friedman, Ori Mosenzon, Noam Slonim, and Naftali Tishby. Multivariate information bottleneck. In *Proceedings of the 17th Conference of Uncertainty in Artificial Intelligence*, 2001.
17. Johannes Fürnkranz. Hyperlink ensembles: A case study in hypertext classification. *Information Fusion*, 3(4):299–312, Dec 2002.
18. R. Ghani. Combining labeled and unlabeled data for multi-class text categorization. In *Proceedings of the International Conference on Machine Learning*, 2002.
19. David Gondek and Thomas Hofmann. Conditional information bottleneck clustering. In *Proceedings of the 3rd IEEE International Conference on Data Mining, Workshop on Clustering Large Data Sets*, 2003.
20. David Gondek and Thomas Hofmann. Non-redundant data clustering. In *Proceedings ot the 4th IEEE International Conference on Data Mining*, 2004.
21. Thorsten Joachims. Transductive inference for text classification using support vector m achines. In *International Conference on Machine Learning (ICML)*, Bled, Slowenien, 1999.
22. Thorsten Joachims. Transductive learning via spectral graph partitioning. In *Proceedings of the International Conference on Machine Learning*, 2003.
23. Karin Kailing, Hans-Peter Kriegel, Alexey Pryakhin, and Matthias Schubert. Clustering multi-represented objects with noise. In *Proceedings og the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2004.
24. S. Kiritchenko and S. Matwin. Email classification with co-training. In *Proceedings of CASCON 2001*, pages 192–201, Toronto, Canada, 2001.
25. Mark-A. Krogel and Tobias Scheffer. Multirelational learning, text mining, and semi-supervised learning for functional genomics. *Machine Learning*, 57(1/2):61–81, 2004.
26. Ion Muslea. *Active learning with multiple views.* PhD thesis, University of Southern California, 2002.
27. K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of Information and Knowledge Management*, 2000.
28. Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.
29. Stefan Rüping. Probabilistic SVMs - how much scaling do we need? Talk at the NIPS 2004 Workshop on Calibration and Probabilistic Prediction in Supervised Learning, Dec. 2004.

30. Stefan Rüping. A simple method for estimating conditional probabilities in SVMs. In A. Abecker, S. Bickel, U. Brefeld, I. Drost, N. Henze, O. Herden, M. Minor, T. Scheffer, L. Stojanovic, and S. Weibelzahl, editors, *LWA 2004 - Lernen - Wissensentdeckung - Adaptivität*. Humboldt-Universität Berlin, 2004.
31. Stefan Rüping. Classification with local models. In Katharina Morik, Jean-Francois Boulicaut, and Arno Siebes, editors, *Proceedings of the Dagstuhl Workshop on Detecting Local Patterns*, Lecture Notes in Computer Science. Springer, 2005. to appear.
32. Matthias Seeger. Learning with labeled and unlabeled data. Technical report, University of Edinburgh, 2000.
33. Ioannis Tsochantaridis and Thomas Hofmann. Support vector machines for polycategorical classification. In *Proceedings of the ECML 2002*, pages 456–467, 2002.
34. Kagan Tumer and Joydeep Ghosh. Order statistics combiners for neural classifiers. In *Proceedings of the World Congress on Neural Networks*, 1995.
35. J. Wang, H. Zeng, Z. Chen, H. Lu, L. Tao, and W. Ma. Recom: Reinforcement clustering of multi-type interrelated data objects. In *Proceedings of the ACM SIGIR Conference on Information Retrieval*, 2003.
36. D. Wolpert. Stacked generalizations. *Neural Networks*, 5:241–259, 1992.
37. D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, 1995.
38. D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, volume 16, pages 321–328, 200a.
39. Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, CMU CALD, 2002.