



UvA-DARE (Digital Academic Repository)

Ensemble-training: ensemble based co-training

Tanha, J.; van Someren, M.; Afsarmanesh, H.

Publication date

2011

Document Version

Final published version

Published in

Benelearn 2011: Proceedings of the Twentieth Belgian Dutch Conference on Machine Learning, The Hague, May 20 2011

[Link to publication](#)

Citation for published version (APA):

Tanha, J., van Someren, M., & Afsarmanesh, H. (2011). Ensemble-training: ensemble based co-training. In P. van der Putten, C. Veenman, J. Vanschoren, M. Israel, & H. Blockeel (Eds.), *Benelearn 2011: Proceedings of the Twentieth Belgian Dutch Conference on Machine Learning, The Hague, May 20 2011* (pp. 123-124). NFI.
http://www.liacs.nl/~putten/benelearn2011/Benelearn2011_Proceedings.pdf

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Benelearn 2011



**Proceedings
of the Twentieth
Belgian Dutch Conference on Machine Learning
The Hague, May 20 2011**

**Editors:
Peter van der Putten
Cor Veenman
Joaquin Vanschoren
Menno Israel
Hendrik Blockeel**

Benelearn 2011

Proceedings
of the Twentieth
Belgian Dutch Conference on Machine Learning
The Hague, May 20 2011

Benelearn 2011. Proceedings of the Twentieth Belgian Dutch Conference on Machine Learning The Hague, May 20 2011.

<http://www.benelearn2011.org>

© 2011 - Benelearn 2011 organization committee and the paper authors(s)/owner(s).

Table of Contents

Preface	9
Organization	11
Invited Talks	13
Kim Luyckx. <i>Text Analytics and Machine Learning for User Identification and Content Reliability</i>	14
Aris Gionis. <i>Efficient Tools for Mining Large Graphs</i>	15
Arno Siebes, <i>Utrecht University, MDL for Pattern Mining</i>	16
Distinguished full papers	17
Bas Weijtens, Gwennyn Kapitein and Marten den Uyl. <i>Geopredict: Exploring Models for Geographical Crime Forecasting</i>	19
Erik Tromp and Mykola Pechenizkiy. <i>Graph-Based N-gram Language Identification on Short Texts</i>	27
Bo Gao and Joaquin Vanschoren. <i>Visualizations of Machine Learning Behavior with Dimensionality Reduction Techniques</i>	35
Full papers	
Bo Gao and Bettina Berendt. <i>Inspecting Patterns for Higher-order Relations: Visualizations for Discrimination-aware and Privacy-aware Data Mining</i>	45
Menno van Zaanen, Tanja Gaustad and Jeanou Feijen. <i>Influence of Size on Pattern-based Sequence Classification</i>	53
Adrian M. Kentsch, Walter A. Kusters, Peter van der Putten and Frank W. Takes. <i>Exploratory Recommendations Using Wikipedia's Linking Structure</i>	61
Arno Knobbe, Pieter Hoogestijn and Durk Kingma. <i>A Feature Construction and Classification Approach to Pixel Annotation</i>	69
Extended Abstracts	
Zoltan Szlavik, Bart Vaerenberg, Wojtek Kowalczyk and Paul Govaerts. <i>Opti-Fox: Towards the Automatic Tuning of Cochlear Implants</i>	79
Frank Takes and Walter Kusters. <i>Determining the Diameter of Small World Networks</i>	81
Pedro Nogueira and Zoltan Szlavik. <i>Automatic Construction of Personalized Concept Maps from Free Text</i>	83

Eelco den Heijer and A.E. Eiben. <i>Multi-Objective Evolutionary Art</i>	85
Vincent Van Asch and Walter Daelemans. <i>Using Domain Similarity for Performance Estimation</i>	87
Zoltan Szlavik, Wojtek Kowalczyk and Martijn Schut. <i>Diversity Measurement of Recommender Systems under Different User Choice Models</i>	89
Dejan Radosavljevik, Peter van der Putten and Kim Kyllesbech Larsen. <i>Customer Satisfaction and Network Experience in Mobile Telecommunications</i>	91
Huu Minh Nguyen, Ivo Couckuyt, Dirk Gorissen, Yvan Saeys, Luc Knockaert and Tom Dhaene. <i>Avoiding overfitting in surrogate modeling: an alternative approach</i>	93
Jan Verwaeren, Willem Waegeman and Bernard de Baets. <i>Incorporating prior knowledge in multiple output regression with application to chemometrics</i>	95
Alexander Nezhinsky, Irene Martorelli and Fons Verbeek. <i>Detection of Developmental Stage in Zebrafish Embryos in a High throughput Processing Environment</i>	97
Sander Bohte. <i>Fractionally Predictive Spiking Neural Networks</i>	99
Nele Verbiest, Chris Cornelis and Francisco Herrera. <i>Granularity based Instance Selection</i>	101
Gilles Louppe and Pierre Geurts. <i>A zealous parallel gradient descent algorithm</i>	103
Eduardo Costa, Celine Vens and Hendrik Blockeel. <i>Protein Subfamily Identification using Clustering Trees</i>	105
Nikolaj Tatti and Boris Cule. <i>Mining Closed Strict Episodes</i>	107
Mihail Mihaylov, Yann-Aël le Borgne, Karl Tuyls and Ann Nowé. <i>DESYDE: Decentralized (De)synchronization in Wireless Sensor Networks</i>	109
Nathalie Jeanray, Raphaël Marée, Benoist Pruvot, Olivier Stern, Pierre Geurts, Louis Wehenkel and Marc Muller,. <i>Phenotype Classification of Zebrafish Embryos by Supervised Learning</i>	111
Olivier Stern, Raphaël Marée, Jessica Aceto, Nathalie Jeanray, Marc Muller, Louis Wehenkel and Pierre Geurts. <i>Zebrafish Skeleton Measurements using Image Analysis and Machine Learning Methods</i>	113
Matteo Gagliolo, Kevin van Vaerenbergh, Abdel Rodríguez, Stijn Goossens, Gregory Pinte and Wim Symens. <i>Policy gradient methods for controlling systems with discrete sensor information</i>	115
Marvin Meeng and Arno Knobbe. <i>Flexible Enrichment with Cortana — Software Demo</i> ...117	
Marc Mertens, Glen Debar, Jonas van den Bergh, Toon Goedeme, Koen Milisen, Jos Tournoy, Jesse Davis, Tom Croonenborghs and Bart Vanrumste. <i>Towards automatic monitoring of activities using contactless sensors</i>	121

Jafar Tanha, Maarten van Someren and Hamideh Afsarmanesh. <i>Ensemble-Training: Ensemble Based Co-Training</i>	123
Gerben K.D. de Vries, Willem Robert van Hage and Maarten van Someren. <i>Comparing Vessel Trajectories using Geographical Domain Knowledge and Alignments</i>	125
Wessel Luijben, Zoltán Szilávik and Daniel Dahlsveen. <i>Scale-Independent Forecasting Performance Comparison</i>	127
Rob Konijn and Wojtek Kowalczyk. <i>Finding Fraud in Health Insurance Data with Two-Layer Outlier Detection Approach</i>	129
Vaisha Bernard and Cor Veenman. <i>The Discovery of Criminal Behavior as a Ranking Problem</i>	131
Guido Demmenie, Jan van den Berg, Menno Israel, Virginia Dignum and Jos Vrancken. <i>(Exceptional) Work Mining: A Stepwise Approach for Extracting Event Logs from Corporate Network Data</i>	133

Ensemble-Training: Ensemble Based Co-Training

Jafar Tanha, Maarten van Someren, and Hamideh Afsarmanesh

J.TANHA,M.W.VANSOMEREN,H.AFSARMANESH@UVA.NL

Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

Keywords: Semi-Supervised Learning, Co-Training, Ensemble Learning

1. Introduction

There are several different methods for semi-supervised learning. Here we consider co-training (Blum & Mitchell, 1998). Classifiers are trained using two views of data, usually subsets of features. Each classifier predicts labels for the unlabeled data including a degree of confidence. Unlabeled instances that are labeled with high confidence by one classifier are used as training data for the other. In this paper, we propose two improvements for co-training. First we consider co-training as an ensemble of N classifiers that are trained in parallel and second we derive a stop-criterion, using a theorem by Angluin and Laird (Angluin & Laird, 1988) that describes the effect of learning from uncertain data.

Two key issues in Co-Training are (1) measuring the confidence in labels that are predicted for the unlabeled data and (2) a criterion for stopping. Co-Training aims at adding a subset of the most confidently predicted labels. At some point labels will be noisy and cause the result of learning to become worse, a form of “overfitting”. Problems (1) and (2) could be solved in an empirical way, using a holdout set of labeled data or some resampling scheme on the labeled dataset but Semi-Supervised Learning is used for learning tasks where labeled data is scarce. We use a theorem from PAC-learning (1) that relates the number of training data to the probability that a consistent hypothesis has an error larger than some threshold for a setting with training data with a certain error in the labels. We use an ensemble of learners for co-training and we use the agreement between the predictions of labels for the unlabeled data to obtain an estimate of the labeling error. Using this we can estimate the effect of learning on the error of the result of adding the new labeled data to the training set. In particular we use a theorem by Angluin and Laird (Angluin & Laird, 1988). If we draw a sequence σ of m data points then

if

$$m \geq \frac{2}{\epsilon^2(1-2\eta)^2} \ln\left(\frac{2N}{\delta}\right) \quad (1)$$

where ϵ is the classification error of the worst remaining candidate hypothesis, η (< 0.5) is an upper bound on the classification noise rate, N is the number of hypothesis, and δ is the confidence, then a hypothesis H_i that minimizes disagreement with σ will have:

$$Pr[d(H_i, H^*) \geq \epsilon] \leq \delta \quad (2)$$

where $d(\cdot)$ is the sum over the probability of elements from the symmetric difference between the two hypothesis sets H_i and H^* . Adding data with labels that have a probability of being incorrect means that m is increased and that η must be adjusted using the probability of an incorrect label for the current labeled set and of the newly labeled set. We fix δ and we assume that N is approximately constant. In that case we can calculate ϵ . The noise rate in this training set can be estimated by:

$$\eta_{i,j} = \frac{\eta_L |L| + \hat{\epsilon}_{i,j} |L_{i,j}|}{|L| + |L_{i,j}|} \quad (3)$$

From this we derive a criterion for whether adding data reduces the error of the result of learning or not. In Ensemble-training each component classifier h_j is first trained on the original labeled data. Next ensembles are built by using all classifiers except one. These ensembles predict the class of the unlabeled data and also the error rate is estimated from the agreement between the classifiers. After that a subset of U , unlabeled data, is selected by ensemble H_p for a classifier that will reduce its error. Here a threshold on the improvement can be used. This is then added to the labeled training data and the estimated error rate of these data is adjusted and used as training data for the “held-out” classifier. The selected unlabeled data for the subset

in each training process is not removed from the unlabeled data U because may be the other component classifiers use them as well. This is repeated until no more data improve the error.

2. Evaluation and Conclusion

We compared our method to Tri-Training, Co-Forest (Li & Zhou, 2007) and Self-Training on a eight datasets from the UCI repository. Ensemble-training gave the best results on four of these. On average the accuracy was 1.63% higher than Co-Forest, which was second best. Though based on an approximation of the error rate, Ensemble-training gives good results compared to other methods.

References

- Angluin, D., & Laird, P. (1988). Learning from noisy examples. *Mach. Learn.*, 2, 343–370.
- Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *Proceedings of the eleventh annual conference on Computational learning theory* (pp. 92–100). New York, NY, USA: ACM.
- Li, M., & Zhou, Z.-H. (2007). Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 37, 1088–1098.