UNIVERSITY OF AMSTERDAM

# 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.

*UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)*

Download date:02 Jul 2022

# 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

# 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 $\sigma$ of $m$ data points then if

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

where $\epsilon$ is the classification error of the worst remaining candidate hypothesis, $\eta$ ($< 0.5$) is an upper bound on the classification noise rate, $N$ is the number of hypothesis, and $\delta$ is the confidence, then a hypothesis $H_i$ that minimizes disagreement with $\sigma$ will have:

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

where $d(,)$ 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 t hat $\eta$ must be adjusted using the probability of an incorrect label for the current labeled set and of the newly labeled set. We fix $\delta$ and we assume that N is approximately constant. In that case we can calculate $\epsilon$. The noise rate in this training set can be estimated by:

$$\eta_{i,j} = \frac{\eta_L|L| + \hat{e}_{i,j}|L_{i,j}|}{|L| + |L_{i,j}|} \tag{3}$$

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.