# USE OF UNLABELED DATA IN SUPERVISED MACHINE LEARNING

*Blaž Novak*
Department of Knowledge Technologies
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
e-mail: blaz.novak@ijs.si

**ABSTRACT**

**In many machine learning problem domains large amounts of data are available but the cost of correctly labeling it prohibits its use. This paper presents a short overview of methods for using a small set of labeled data together with a large supplementary unlabeled dataset in order to learn a better hypothesis than just by using the labeled information.**

## 1 INTRODUCTION

In the recent years, enormous amounts of information has become available – most notably unstructured and semi-structured textual data available from the internet. In order for this information to be of greater use, more structure needs to be discovered in it – to enable automated processing and reasoning. One of the tools used for this is machine learning.

Supervised machine learning is a process of learning a function based on given examples. The examples are provided as ordered pairs of objects (A, B) and the learning algorithm induces the function f: A → B based on some inductive bias (prior knowledge / assumptions) which is needed for any generalization to be possible. The resulting function can then be used to map objects into unknown target values.

Since the data available can have a large complexity this inherently means complex functions to be learned – and learning complex functions requires many examples. Examples with known target values (a.k.a. labeled examples) are however usually not directly available and need to be manually created, which can be a time-consuming and/or expensive process. In order to minimize this cost, a lot of research has been conducted in the area of using unlabeled examples to aid in the process.

Two different approaches and their mixtures will be presented here; one designed to minimize required human effort and the other to work with a fixed set of labeled and unlabeled examples.

## 2 ACTIVE LEARNING

Active learning is also known as 'experiment design' in statistical literature. In contrast to normal (passive) machine learning where the learner is presented with a static set of examples that are then used to construct a model, active learning paradigm means the learner can 'ask' the oracle/domain expert/user/... for a label of an example. The intuition is that a few highly informative examples provide much more information than a lot of random ones. However, one must be careful since this violates the assumption of randomly sampled input made by a lot of the algorithms. Since in many practical cases construction of queries is hard (e.g. a construction of a meaningful document (which is to be labeled by an expert) from a bag of words model commonly used for document classification is close to impossible) a 'query filtering' [1] paradigm becomes useful: the learner is provided with a large amount of non-labeled examples that are potential queries. It is then its job to select the potentially interesting ones. Since the problem of finding the optimum subset of the most interesting questions is hard, a greedy approximation is used. The basic active learning algorithm is then as follows:

```
start with a small labeled set and a large unlabeled set
repeat until some condition is met:
        from the unlabeled set select the currently most interesting example
        query the expert/oracle/.. for the label
        add the now-labeled example to the labeled set
```

Algorithm 1

The core of research of active learning algorithms is obviously the selection of the most interesting example. At the top level there are two different approaches: indirect and direct classifier optimization.

### 2.1 Indirect

Indirect methods are based on the idea of version space minimization: the size of the set of all possible hypotheses that are still consistent with all of the examples seen so far should be minimized as fast as possible [2]. There exist theoretical proofs of exponential reduction of the number of

required examples under certain assumptions [3], but the most limiting assumption is that there is no noise present in the data.

These approaches can again be divided to single- and multi-model based. With single-model methods [4, 1] an assumption is made that a high certainty prediction by a single model also means that a large portion of models from the current version space would give the same prediction - meaning that after inclusion of that example into the labeled learning set only a small amount of hypotheses would be removed, therefore making that example inappropriate if the goal is to minimize VS as quickly as possible. Examples with *low* prediction certainty are then presented to the oracle.

```
example selection step:

using currently labeled examples train a model that can output a prediction certainty
(e.g. naïve bayes)
for each unlabeled example still available
        predict the target value and remember the classifier certainty
pick N examples with the lowest certainty and submit them for classification
```

Algorithm 2

The obvious problem with this method is that the aforementioned assumption is generally not true. On the positive side the algorithm is relatively fast compared to other AL algorithms.

An extension of this idea is an SVM-specific algorithm [5]: for each example two different models are built – one with the example temporarily put in the positive class and the other with example in the negative class. The example with the most similar margin sizes is selected for labeling. This method however has a large time complexity: for each example considered, a couple of SVM models must be *created* as opposed to just evaluated in the previous methods. A simplification that selects the example only based on the distance from the margin is possible but is similar to the previously mentioned uncertainty sampling algorithms.

The multi-model approach is based on the idea that if we had infinitely many models randomly sampled from the version space we should select the example with the highest prediction entropy considering the sampled models. Such an example will on average remove the largest possible portion of the version space after being labeled. The idea is called "query by committee" or QBC [2].

### 2.2 Direct optimization

The direct approach does not make the assumption that the data is noise-free. It does not try to minimize the version space but instead directly minimizes the expected future prediction error of the final model over the entire sample space and so directly optimizes the criteria function with which the model will be evaluated. At each step such an example is selected that would – if added labeled – minimize the expected error. Since data needed to estimate that is not known, an approximation is again used. One possible approximation is to select such an example that results in a model with the minimum average prediction variance over the unlabeled set if it is added into the labeled set with all the possible target values [6]. Another option is to minimize the expected loss over a validation subset generated from the existing labeled data [7, 23].

### 2.3 Summary

The performance of the aforementioned techniques (measured by the expert labeling cost) can vary from problem to problem by orders of magnitude. Computational cost should also be taken into account when choosing an approach - while decreasing the cost of human labor the CPU requirements can increase beyond any reasonable limit: in the usual learning scenario one only needs to train one model which can already be an expensive procedure. For a simple uncertainty-based active learning, one has to train the same number of models as there are labeled examples at the end and use every one of them to test each unlabeled sample. It is possible to decrease the amount of CPU work by a constant factor at the expense of some human labor by selecting several examples at the same time without updating the rest of the system. For the method based on SVM margin sizes, the number of trained and discarded models is for each iteration of active learning loop linearly dependant on the size of the unlabeled set; making efficient implementation of incremental learning algorithms an absolute must.

## 3 SEMI-SUPERVISED LEARNING

While also dealing with unlabeled data, semi-supervised learning [8] is not an interactive procedure. The algorithm is provided with a set of labeled examples and a set of unlabeled examples which can be used as an addition to gain an insight on the data.

One possible use of unlabeled examples is to correct the sampling bias if the labeled examples have been sampled nonrandomly [9]. Otherwise the labeled data can either improve or decrease the models accuracy – depending on the distribution and model assumptions.

### 3.1 Semi-supervised transduction

The first possibility in semi-supervised learning is transduction: one only has to label the already known unlabeled data. One common approach is to first construct a graph using all of the examples as vertices and connect those vertices that are similar – close to each other according to a chosen distance measure – and assign that distance as the weight of the edge. Labels from the labeled examples are then propagated to the unlabeled ones.

The simplest algorithm for assignment of binary labels is based on graph mincut [10]:

construct a graph using all of the examples
add two more vertices (one for each label) (+), (-)
connect labeled vertices with the corresponding (+) or (-) vertex with edges of infinite weight
connect the rest of the vertices with edges weighted by the similarity function
find the minimum cut between (+) and (-) thus minimizing the number of similar vertices that will be given different labels
assign labels to the unlabeled vertices depending on which side of the cut they are

Algorithm 3

Since there is a possibility for these cuts to be degenerated (e.g. if the graph is a path with all of the edge weights equal there are n-1 possible cuts but the mincut algorithm will return one of the cuts with one vertex on one side and the rest on the other – which is clearly not a desirable solution) other possibilities of assignment exist – from randomized version of this algorithm [11] to using spectral graph partitioning [12], random walks [13] and Gaussian random fields with respect to the weighted graph [14].

### 3.2 Semi-supervised induction

Semi-supervised inductive methods are mostly based on expectation maximization (EM) – an iterative algorithm for improving the hypothesis. The general idea is to create the initial model, label the unlabeled data and then iteratively generate a new model using all of the labels, relabel the originally unlabeled data using that model; stopping when some convergence criteria is met. The problem with this approach is that a lot of incorrectly labeled examples in the initial steps of the algorithm can mask the labeled examples, forcing the model to converge to a random point [8]. A weighting of the samples must therefore be used. Alternative solutions have also been proposed [16].

In the case that multiple independent views (i.e. two or more *independent* sets of attributes describing the same examples) of data are available maximization of prediction consistency across models trained on different views can be attempted. The 'co-training' [17] algorithm iteratively learns multiple models (one on each view) and allows each of them to label some unlabeled examples. The examples with the most confident prediction are then added to the labeled set and the process is repeated.

The Co-EM [18] algorithm combines EM and co-training. It uses the hypotheses learned from one view to probabilistically label the examples which are then used to learn a hypothesis on another view.

### 3.3. Constrained clustering

Constrained clustering is clustering with background knowledge. Constraints can be instance based [19] (e.g. two examples must / must not be in the same cluster), hard (mandatory) or soft (unobserved constraints add penalty), global constraints [20] (e.g. each cluster must have at least N elements) or even in a form of declarative knowledge (a subset of FOL in [21]). Conversion of ordinary clustering algorithms into constrained clustering is quite straightforward. The constraints even reduce the size of the search space, making the algorithms faster than their counterparts without constraints.

## 4 MIXTURES

Active learning and semi-supervised learning can also be merged into one process. If multiple views on the data are available, one can learn multiple hypotheses and choose for labeling those examples on which the most of the hypotheses disagree – the algorithm being called Co-testing.

Co-testing and Co-EM can also be interleaved into Co-EMT [22]: hypotheses for co-testing are learned by Co-EM algorithm on both the labeled and unlabeled examples.

## 5 CONCLUSION

In this paper a short overview of the possibilities of unlabeled data's contribution to a learning task has been given. There are still a lot of open questions about the actual theoretical value of such information but it seems that in practice they significantly improve the results.

## REFERENCES


[1] "A sequential algorithm for training text classifiers", David D. Lewis and William A. Gale. Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval.

[2] "Query by Committee", H. S. Seung and Manfred Opper and Haim Sompolinsky. Computational Learning Theory pg. 287-294, 1992.

[3] "Information, prediction, and query by committee", Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Advances in Neural Information Processing Systems 5, pages 483-490, 1993.

[4] "Improving Generalization with Active Learning", David A. Cohn and Les Atlas and Richard E. Ladner. Machine Learning 15-2, pg. 201-221, 1994.

[5] "Support Vector Machine Active Learning with Applications to Text Classification", Simon Tong and Daphne Koller. Proceedings of ICML-00, 17th International Conference on Machine Learn ing, pg. 999-1006.

[6] "Active Learning with Statistical Models", David A. Cohn and Zoubin Ghahramani and Michael I. Jordan. Advances in Neural Information Processing Systems vol 7, pg. 705-712, 1995.

[7] "Toward Optimal Active Learning through Sampling Estimation of Error Reduction", N. Roy, A. McCallum.

[8] "Learning with Labeled and Unlabeled Data", Matthias Seeger. Technical Report, Edinburgh University.

[9] "Automatic Bayes Carpentry Using Unlabeled Data in Semi-Supervised Classification", H. Zou, J. Zhu, T. Hastie.

[10] "Learning from Labeled and Unlabeled Data Using Graph Mincuts", Avrim Blum and Shuchi Chawla. Proc. 18th International Conf. on Machine Learning, pg. 19-26, 2001.

[11] "Semi-Supervised Learning Using Randomized Mincuts", A. Blum, J. Lafferty, M. R. Rwebangira, R. Reddy. In Proceedings of the 21st International Conference on Machine Learning, 2004.

[12] "Transductive Learning via Spectral Graph Partitioning", Thorsten Joachims. In Proceedings of the Twentieth International Conference on Machine Learning, 2003.

[13] "Learning from Labeled and Unlabeled Data Using Random Walks", D. Zhou, B. Scholkopf.

[14] "Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions", Xiaojin Zhu, Zoubin Ghahramani, John Lafferty. In Proceedings of The Twentieth International Conference on Machine Learning, 2003.

[16] "Stable Mixing of Complete and Incomplete Information", Adrian Corduneanu and Tommi Jaakkola. AI Memo 2001-030.

[17] "Combining Labeled and Unlabeled Data with Co-training", Avrim Blum and Tom Mitchell. COLT: Proceedings of the Workshop on Computational Learning Theory, 1998.

[18] "Analyzing the Effectiveness and Applicability of Co-training", Kamal Nigam and Rayid Ghani.

[19] "Clustering with Instance-level Constraints", K. Wagstaff, C. Cardie.

[20] "Constrained K-Means Clustering", P. S. Bradley, K. P. Bennett, A. Demiriz. MSR-TR-2000-65.

[21] "Integrating Declarative Knowledge in Hierarchical Clustering Tasks", L. Talavera and J. Bejar.

[22] "Active + Semi-Supervised Learning = Robust Multi-View Learning", I. Muslea and S. Minton and C. A. Knoblock.

[23] "Active Learning for Class Probability Estimation and Ranking", M. Saar-Tsechansky and F. Provost.