In *Proceedings of the Sixth International Colloquium on Cognitive Science*,
San Sebastian, Spain, 1999 (invited paper).

The Role of Unlabeled Data in Supervised Learning

Tom M. Mitchell
School of Computer Science
Carnegie Mellon University

Tom.Mitchell@cmu.edu

ICCS-99
May 1999

1 Introduction

Most computational models of supervised learning rely only on labeled training examples, and ignore the possible role of unlabeled data. This is true both for cognitive science models of learning such as SOAR [Newell 1990] and ACT-R [Anderson, et al. 1995], and for machine learning and data mining algorithms such as decision tree learning and inductive logic programming (see, e.g., [Mitchell 1997]). In this paper we consider the potential role of *unlabeled* data in supervised learning. We present an algorithm and experimental results demonstrating that unlabeled data can significantly improve learning accuracy in certain practical problems. We then identify the abstract problem structure that enables the algorithm to successfully utilize this unlabeled data, and prove that unlabeled data will boost learning accuracy for problems in this class. The problem class we identify includes problems where the features describing the examples are redundantly sufficient for classifying the example; a notion we make precise in the paper. This problem class includes many natural learning problems faced by humans, such as learning a semantic lexicon over noun phrases in natural language, and learning to recognize objects from multiple sensor inputs. We argue that models of human and animal learning should consider more strongly the potential role of unlabeled data, and that many natural learning problems fit the class we identify.

2 A Supervised Learning Problem: Learning to Classify Web Pages

To illustrate the role of unlabeled data in supervised learning, consider the problem of learning to classify pages of hypertext from the world wide web, given labeled training data consisting of individual web pages along with their correct classifications (Figure 1). We have been studying this problem as part of our larger research goal of automatically extracting information from the web [Craven et al.

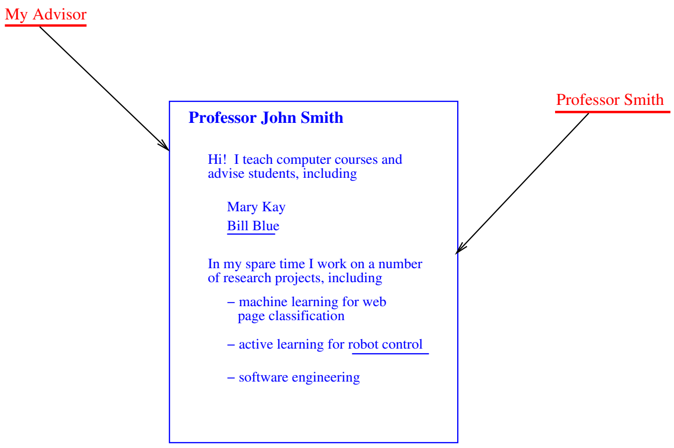

Figure 1: Training example of a "faculty home page." The task of classifying a web page can be achieved by considering just the words on the web page. Alternatively, the page can be classified using only the words on hyperlinks that point to the web page (e.g., "my advisor" "Professor Smith"). When examples are described by such *redundantly sufficient features*, unlabeled data can be used to boost the accuracy for supervised learning.

1998]. For example, one specific task we have considered is training a system to classify web pages into categories such as "student home page," "faculty home page," "home page of an academic course," etc.

This web page classification problem can be formulated as a typical supervised learning problem. By "supervised learning problem" we mean that there is some set X of instances, and there is some target function $f : X \rightarrow Y$ to be learned, given a set of training examples of the form $\{\langle x_i, f(x_i)\rangle\}$. In our case, X is the set of all web pages, Y is the set of possible classifications (e.g., "student" "faculty" "neither"), and the function f to be learned in the function that maps an arbitrary page x to its correct classification. For the ith labeled training example, x_i is a particular web page, and $f(x_i)$ is the correct classification for x_i, provided by an external teacher.

As discussed in [Craven et al. 1998], a number of supervised learning algorithms can be applied to the problem of learning to classify web pages. One common approach is to first represent each web page by a large feature vector, where each possible word in the language corresponds to a feature, and the value of the feature is the number of times the word occurs in the web page. This "bag of words" representation obviously ignores information about the sequence in which words occur. However, this representation of web pages as large feature vectors enables us to apply standard supervised learning algorithms for classification of feature vectors. For example, if we train a naive Bayes classifier [Mitchell 1997] on a set of approximately 4,000 labeled web pages, we achieve accuracies of approximately 70% in classifying new web pages into the categories mentioned above [Craven et al., 1997].

Given:

- set L of labeled training examples
- set U of unlabeled examples

Loop:

- Learn hyperlink-based classifier H from L
- Learn full-text classifier F from L
- Allow H to label p positive and n negative examples from U
- Allow F to label p positive and n negative examples from U
- Add these self-labeled examples to L

Table 1: A Cotraining algorithm, for training two classifiers and using unlabeled data to boost their accuracies.

While we can achieve reasonable accuracy after training on thousands of hand labeled web pages, in this paper we are interested in the question of how unlabeled data can be useful. In fact, this is a crucial question in our web page classification task, because it is costly to hand label thousands of web pages, and because it is very easy to obtain hundreds of millions of unlabeled pages from the web.

How can we use unlabeled web pages to more accurately learn to classify future pages? The answer can be seen by first observing that the web contains an interesting kind of redundant information about each web page, as shown in figure 1. As shown in this figure, we could classify the web page either by considering the words on the page itself (as suggested in the paragraph above), or we could classify the page by ignoring the words on the page and instead considering the words on hyperlinks that point to the page. In many cases, the words on the hyperlinks will be sufficient to classify the example, and the words on the page itself will also be sufficient. In such cases, we will say these two sets of features – the hyperlink words, and the web page words – are *redundantly sufficient* to classify the example. We will see below that in general if the instances X for some supervised learning task can be factored into sets of redundantly sufficient features, then unlabeled data can be used to boost the accuracy of classifiers trained with limited labeled data.

3 Learning with Redundantly Sufficient Features

How can we use unlabeled data to boost the learning accuracy for our web classification problem?

	Page-based classifier	Hyperlink-based classifier	Combined classifier
Supervised training	12.9	12.4	11.1
Co-training	6.2	11.6	5.0

Table 2: Error rate in percent for classifying web pages as course home pages. The top row shows errors when training on only the labeled examples. Bottom row shows errors when co-training, using both labeled and unlabeled examples. The rightmost column shows the classification accuracy when the two classifiers are provided equal votes on the final classification. The results reported here are the average of five independent runs of the algorithm with different random starting examples.

The key idea is that we will train two independent classifiers rather than one. One classifier will use only the words on the web page, and the other classifier will use only the words on the hyperlinks. We begin by training both classifiers using whatever labeled training examples are available. Presumably this will result in two classifiers that are imperfect, but better than random. Now we use the unlabeled data as follows: each classifier is allowed to examine the unlabeled data and to pick its most confidently predicted positive and negative examples, and add these to the set of labeled examples. In other words, each classifier is allowed to augment the pool of labeled examples. Both classifiers are now retrained on this augmented set of labeled examples, and the process is repeated as long as desired. Table 1 summarizes this co-training algorithm.

Why should this co-training algorithm lead to more accurate classifiers? The intuition is that if the hyperlink classifier finds an "easily classified" hyperlink in the unlabeled data (e.g., one that is quite similar to one of the labeled examples on which it was trained), the web page that it points to will be added to the labeled pool of examples as well. Of course just because the hyperlink happened to be easy to classify does not mean the web page will be easily classified by the other classifier. If not, then the hyperlink classifier has added useful training information to improve the other classifier. Similarly, the web page classifier can add examples that are easy for it to classify, but that provide useful information to improve the accuracy of the hyperlink classifier.

In experiments, we have found that this Cotraining algorithm does improve classification accuracy when learning to classify web pages. In one experiment [Blum and Mitchell, 1998], summarized in Table 2, we trained a classifier to label web pages as home pages of academic courses. In this experiment we provided just 16 labeled examples, and approximately 800 unlabeled pages drawn from computer science department web sites. On each iteration of co-training, each classifier was allowed to add 1 new positive and 3 new negative examples to the pool of labeled examples. After 30 iterations of the cotraining algorithm, the accuracy of the combined classifier was 95%, compared to 89% when only the labeled data was used. In this case, the impact of cotraining was to reduce the error by better than a factor of two.

4 Formal Results

Given the experimental evidence that cotraining can be useful in at least one case, it is useful to characterize the general problem setting in which this kind of algorithm can benefit from unlabeled data. Here we define a problem setting in which we can prove that unlabeled data will be of help.

To define the general problem setting that captures the essential structure of our web classification example, recall that earlier we defined a supervised learning problem as one where we have some set of instances X, a target function $f : X \rightarrow Y$, and labeled examples $\{\langle x_i, f(x_i)\rangle\}$. In the *cotraining* setting, we assume such a supervised learning problem in which the instances are drawn from X according to some fixed (possibly unknown) probability distribution $\mathcal{D}$. We further assume that $X = X_1 \times X_2$; that is, the instances in X can be factored into two parts (e.g,. the hyperlink words and the web page words). We require also that X_1 and X_2 each contain information sufficient to classify the example. In other words, we require that there exist some function $g_1 : X_1 \rightarrow Y$ and some function $g_2 : X_2 \rightarrow Y$ such that for all $x \in X, g_1(x_1) = g_2(x_2) = f(x)$. Note the learner is not expected to know f, g_1 or g_2 in advance. We simply require that it be possible to express f in terms of x_1 and in terms of x_2 (i.e., that X_1 and X_2 both be sufficient to classify X according to f). If the above constraints on X_1 and X_2 are satisfied, then we will say that X_1 and X_2 are *redundantly sufficient* to classify X with respect to f.

In [Blum and Mitchell 1998] we show that if $f : X \rightarrow Y$ is PAC learnable from noisy labeled data, X_1 and X_2 are redundantly sufficient to classify X with respect to f, and if x_{B_1} and x_{B_2} are conditionally independent given $f(x)$, then f can be PAC learned given a weak initial classifier plus only *unlabeled* data. What this means is that if the labeled data are sufficient to train a better than random initial classifier, if X_1 and X_2 are distributed independently for each target value of f, and if f can be learned accurately given an arbitrary number of *labeled* examples, then f can be learned to similar accuracy given just this weak initial classifier and an arbitrary number of *unlabeled* examples. The significance is that this proves that unlabeled examples can substitute for labeled examples, under the conditions of the theorem.

In [Blum and Mitchell 1998] we also show that unlabeled data is of potential value under the more general setting in which X_1 and X_2 are not independently distributed. There we show that if the training data is noise-free, then unless X_1 and X_2 are deterministically related (e.g., one value of x_1 is always paired with the same x_2), the unlabeled data can be of use.

5 Other Learning Tasks with Redundantly Sufficient Features

Given the success of cotraining as a method for using unlabeled data when learning to classify web pages, and given the formal characterization of the class of problems for which cotraining can be of use, it is interesting to ask what other natural learning problems allow this approach. Here we summarize a number of tasks:

- Learning to classify noun phrases into semantic classes. In [Riloff and Jones 1999], an approach similar to cotraining was applied to the problem of learning a semantic lexicon over noun phrases in English. In this paper, one task was to learn to classify noun phrases as positive or negative examples of locations (e.g., "San Sebastian" is positive, "blue" is negative). In this case each example, x, is a sentence such as "We are located in lovely Pittsburgh." The factorization of x into two redundantly sufficient feature sets is done as follows: x_1 is the noun phrase itself (e.g., "Pittsburgh" in the above sentence), and x_2 is the linguistic context in which the noun phrase appears (e.g., "We are located in lovely —"). Note that in most cases it is possible to determine whether the noun phrase is a location given either x_1 or x_2. Note also that the values of x_1 and x_2 are distributed fairly independently (i.e., with might see the same x_2 as above, but with a different x_1). Starting with a list of just 10 known noun phrases, their system used an algorithm similar to cotraining to learn dozens of additional locations.

- Learning to select word sense. In [Yarowsky, 1995], an approach similar to cotraining was applied to learning to disambiguate word senses (e.g., to determine whether the word "plant" refers to a manufacturing plant or to a botanical plant). Here, each instance x corresponds to a context containing the word in question (similar to the Riloff and Jones example above). However, the factorization of x is into more than two components, with each word serving as a candidate component, and this algorithm seems to not map directly into cotraining. But it is quite similar, and the problem could be attacked using a cotraining approach.

- Learning to recognize phonemes in speech. In [de Sa and Ballard, 1998], an approach to fully unsupervised learning was applied to a problem with nearly redundantly sufficient features. In this case, the task is to learn to classify speech phonemes, based on both the audio signal and the video signal watching the speaker's lips. Here each instance x corresponds to the full data, x_1 corresponds to the audio signal, and x_2 corresponds to the video signal. Again, we may roughly assume that either of x_1 or x_2 is sufficient to determine the phoneme, and therefore can train two classifiers on limited labeled data, and let them train each other over the unlabeled data. In fact, de Sa and Ballard use no labeled data at all, instead performing a kind of clustering in which the audio signal must predict the video, and vice versa. Their system was able to learn clusters corresponding to the spoken phonemes in the data.

- Object recognition in multimedia data. Consider the problem of learning to recognize objects in multimedia data, such as the continuous stream of audio, video, and other sensory input a person receives. A very similar task is to learn to classify television segments, based on the ongoing stream of video, audio, and close captioned text. In this latter case, consider the problem of learning to spot television segments in which Boris Yeltsin appears. Here we could consider each instance x to be a snapshot containing the audio, video and text at a particular time. Here x_1 could be the audio, x_2 the video, and x_3 the text. Note that in some cases we might see "easy" example of video (a full face of Yeltsin), or an easy audio (his voice without background noise), or an easy text (the word Yeltsin). Thus, we could expect cotraining to provide a useful approach to using unlabeled data to train classifiers. To my knowledge, this experiment has not been attempted.

6 Discussion and Conclusions

We have described a class of supervised learning problems for which unlabeled data can be proven to improve learning accuracy. The key defining features of this problem class are that (1) the instances X can be factored into two or more components X_1, $X_2 \ldots$, which are *redundantly sufficient* to classify the example, and (2) these components covary, so that a particular x_i does not always co-occur with the same x_j..

We have described the cotraining algorithm that uses unlabeled data for such problems, and have presented experimental and theoretical results showing that in these problems unlabeled data can indeed be useful to improve accuracy for supervised learning.

The connection to human and animal learning appears to be a rich area for future study. Humans and other animals have a rich set of sensory input, which includes redundantly sufficient data for many tasks. For example, smell, vision, sound can all be useful in trying to classify whether food is nearby, and in many cases the data from just one of these channels is sufficient. de Sa and Ballard's success with a cotraining-like algorithm for learning to classify phonemes from video and audio – similar to that observed by humans – is also suggestive. Perhaps people rely less than we suspect on labeled data to achieve successful learning for various classification tasks.

7 Acknowledgements

Thanks to my collaborator on this work, Avrim Blum, for many interesting discussions and ideas. Sebastian Thrun, Rosie Jones, Andrew McCallum, and Kamal Nigam have also contributed ideas and suggestions along the way. This work has been supported by Darpa under research contract F30602-97-1-0215.

8 References

[Anderson et al. 1995] Production system models of complex cognition. In Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society [pp. 9-12]. Hillsdale, NJ: Lawrence Erlbaum Associates.

[Blum and Mitchell, 1998] Combining Labeled and Unlabeled Data with Co-Training, COLT98.

Available at http://www.cs.cmu.edu/∼webkb.

[Craven et al., 1998] Learning to extract symbolic knowledge from the world wide web. in *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98).* Available at http://www.cs.cmu.edu/∼web

[de Sa, 1994] Learning classification with unlabeled data, NIPS-6, 1994.

[de Sa and Ballard, 1998] Category learning through multi-modality sensing, *Neural Computation* 10(5), 1998.

[Riloff and Jones, 1999] Learning dictionaries for information extraction by multi-level bootstrapping, *AAAI99*, to appear. Available at http://www.cs.cmu.edu/∼webkb.

[Mitchell 1997]. *Machine learning.* New York: McGraw Hill, 1997. See http://www.cs.cmu.edu/∼tom/mlbook.htm

[Newell 1990]. *Unified theories of cognition.* Cambridge, MA: Harvard University Press, 1990.

[Yarowsky, 1995] Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the 33rd Annual Meeting of the ACL*, pp. 189-196.