

TAKING ADVANTAGE OF UNLABELED DATA WITH THE ORDERED CLASSIFICATION ALGORITHM

Thamar Solorio and Olac Fuentes
Computer Science Department
National Institute of Astrophysics, Optics and Electronics
Luis Enrique Erro #1
Tonantzintla, 72840 Puebla, México
email: thamy@cseg.inaoep.mx, fuentes@inaoep.mx

ABSTRACT

We introduce a new method for improving poor performance of classifiers due to a small training set. The Ordered Classification algorithm presented here incrementally increases the training set by adding unlabeled examples. These unlabeled examples are selected by the algorithm accordingly to the confidence level of the predictions made by an ensemble of classifiers. The use of this confidence level measurement, which was inspired by the Query By Committee approach within the Active Learning setting, ensures that the algorithm incorporates the examples which are more likely to have the right classification label assigned by the ensemble. Experimental results show that this algorithm effectively takes advantage of the unlabeled data yielding an error reduction of up to 78%. Giving that a very common scenario in classification problems is the lack of a large enough training set, this algorithm provides a practical solution.

KEY WORDS

machine learning, unlabeled data, ensembles

1 Introduction

In the last few years we have seen in the machine learning community an increasing interest in the use of unlabeled data to aid classifiers. Several approaches, including our own work, have been proposed, proving that unlabeled data can help the classifier to build a better hypothesis about the target function when the labeled examples are scarce. The time and cost needed to build a large enough training set for use with traditional learning algorithms can be reduced considerably if we take advantage of the unlabeled examples using new methodologies. Moreover, we can diminish the risk of having an inaccurate training set if we decrease the number of manually labeled examples needed for training: it has been observed that, due to fatigue the accuracy of manual labelling generally decreases as the number of examples to be labeled increases. Given the high cost of manually labeling examples, the use of unlabeled data represents a real advantage when it comes to building accurate classifiers.

One key issue that has to be solved when we want to increase classifier accuracy with unlabeled data is that of which unlabeled examples to use. Previous works have shown that in some cases the unlabeled data do not help, and may even degrade, the classifier performance (e.g. Cozman and Cohen [1], Nigam et al. [2]). However, we believe that this fact can be avoided with a good discriminative selection of unlabeled examples.

In this paper we introduce an algorithm called Ordered Classification (OC), which uses an ensemble of classifiers to select the unlabeled examples to be added in the training process. By adding these previously unlabeled examples, the learning algorithm is provided with a larger training set, which allows a better approximation of the target function. Classification with the OC is performed by a discriminant approach similar to that of Query By Committee within the active learning setting [3, 4, 5]. The selection criterion is given by the entropy on the predictions made by the members of an ensemble of classifiers. The entropy is considered as our measurement of the confidence level in the predictions made by the ensemble. According to this, examples with lower entropy are more likely to have the correct label assigned by the ensemble than those with higher entropy.

The OC algorithm is an iterative process, in each iteration the algorithm adds to the training set those examples for which the ensemble can assign labels with high confidence, and the process is repeated until there are no unlabeled examples left. The base learning algorithm used here is C4.5 [6]. Experimental results of applying our algorithm to several learning tasks taken from the UCI Machine Learning Repository [7] demonstrate that the use of such selection criterion is a practical solution that can help decrease the classification error by up to 78%.

In previous work [8], we applied the idea of OC to the prediction of stellar atmospheric parameters based on spectral information. As opposed to the learning tasks presented here, in [8] the target function was real valued so the confidence level for selecting unlabeled examples was based on the standard deviation of the ensemble predictions. The base learning algorithm used was Locally Weighted Linear Regression [9] and we attained an error reduction of 29%.

The remainder of this paper is organized as follows: the next section describes the most recent related work organized by idea. Section 3 introduces our algorithm, while in Section 4 we give a brief overview of the base learning algorithm, C4.5. Experimental results are discussed in Section 5. Finally, some conclusions and directions for future work are presented in Section 6.

2 Related Work

Among the most widely used approaches targeted to the use of unlabeled data are the ones based on a generative model. The Expectation Maximization (EM) algorithm in combination with the Naive Bayes classifier is one of them (e.g. Nigam et al. [2]). The EM algorithm iteratively assigns the most probable labels to the unlabeled data and a more probable model is built using this pseudo-labeled data. While this approach has proven to increase classifiers accuracy when added unlabeled data in some problem domains, it is not always applicable. If strong disagreement between the model assumed by the classifier and the real parametric model that generated the data is present, unlabeled data will deteriorate accuracy [1, 10]. This method has been used in text classification domains (e.g. Nigam et al. [2], Nigam [11]), face pose determination (e.g. Baluja, [12]) and remote sensing (e.g. Shahshahani and Landgrebe [13]), and in almost all these domains deteriorations due to unlabeled data is present in some experimental results.

A different approach for using unlabeled data to improve classifier accuracy is the co-training strategy. In [14] Blum and Mitchell presented a co-training algorithm that consists of using two classifiers built with different subsets of attributes, which are assumed to be redundant. The training set is augmented with the examples labeled by the two classifiers, and this allows each classifier to discover useful information about the target function with the examples labeled by the other classifier. A shortcoming of this method is that not all classification problems have two redundant subsets of attributes that are enough for perfect classification. Goldman and Zhou [15] presented a new co-training strategy that improves the performance of standard supervised learning algorithms. Unlike the Blum and Mitchell procedure, they used two different algorithms for bootstrapping from the unlabeled data, ID3 (Quinlan, [16]) and HOODG (Kohavi, [17]). This approach, as well as Blum and Mitchell’s co-training strategy, was applied to text classification problems.

Other proposals for the use of unlabeled data include the use of neural networks [18], graph mincuts [19], Semi-Supervised Support Vector Machines [20] and Kernel Expansions [21], among others.

3 The Ordered Classification Algorithm

The goal of the OC algorithm is to select those examples whose class can be predicted by the ensemble with a high

confidence level in order to use them to improve its learning process by gradually augmenting an originally small training set.

Inspired by the measurement of agreement among committee members used in previous works within the active learning setting (e.g. [3, 4, 5]) the OC algorithm uses the entropy as a measure of the confidence level. Given an ensemble of k members built with the training data available, the confidence level of an example x can be computed according to the following equation:

$$A(x) = \left(\sum_i -p_i \log_2 p_i \right)^{-1} \quad (1)$$

Where p_i is defined as the proportion of classifiers that voted for class i in k , and $|i|$ =number of possible classes. Once the ensemble has evaluated all the unlabeled examples available, the OC algorithm selects the n examples that presented the highest confidence level, as defined in the equation above.

Our algorithm proceeds as follows: First, we build several classifiers (the ensemble) using the base learning algorithm. In this work each classifier that makes up the ensemble is built using the C4.5 (Quinlan [6]) learning algorithm (see the next section for a short overview of this algorithm). For constructing the ensemble we used the technique called *bagging* (Breiman [22]). In this technique, each member of the ensemble has a training set consisting of m examples selected randomly with replacement from the original training set of m examples (Dietterich [23]). Then, each single classifier predicts the classes for the unlabeled data and we use those predictions to measure the confidence level for each example (Equation 1). We now proceed to select the n previously unlabeled examples with the highest confidence level and add them to the training set. Also, the ensemble re-classifies all the examples added until then, if the confidence level is higher than the previous value then the labels of the examples are changed. This process is repeated until there are no unlabeled examples left. See Table 1 for an outline of our algorithm.

4 The Base Learning Algorithm C4.5

C4.5 is an extension to the decision-tree learning algorithm ID3 [16]. Only a brief description of the method is given here, more information can be found in [6]. The algorithm consists of the following steps:

1. Build the decision tree from the training set (conventional ID3).
2. Convert the resulting tree into an equivalent set of rules. The number of rules is equivalent to the number of possible paths from the root to a leaf node.
3. Prune each rule by removing any preconditions that result in improving its accuracy, according to a validation set.

I_s is a matrix whose rows are vectors of attribute values
 L_s is the class label
 S is the training set, given by the tuple $[I_s, L_s]$
 U is the unlabeled test set
 A is initially empty and will contain the unlabeled examples added to the training set

1. While $U \neq \emptyset$ do:
 - Construct E , the ensemble containing k classifiers
 - Classify U and estimate reliability of predictions
 - V are the n elements of U for which the classification assigned by the ensemble is most reliable:
 - $S = S \cup V$
 - $U = U - V$
 - $A = A \cup V$
 - Classify A using E and change the labels of the examples with higher confidence level
2. End

Table 1. The algorithm of ordered classification

4. Sort the pruned rules in descending order according to their accuracy, and consider them in this sequence when classifying subsequent instances.

Since the learning tasks used to evaluate this work involve nominal and numeric values, we implemented the version of C4.5 that incorporates continuous values.

5 Experimental Results

In the experiments performed in this work we used the evaluation technique 5-fold cross-validation, which consists of randomly dividing the data into 5 equally-sized subgroups and performing 5 different experiments. We separated one group along with their original labels as the validation set; another group was considered as the test set; from the remaining data we randomly selected 10 examples with their original labels as the starting training set and the remainder of the data was considered the unlabeled set. Each experiment consists of ten runs of the procedure described above, and the overall average are the results reported here. The learning tasks used here are: balloons, lymphography, breast cancer, wine, tic-tac-toe and iris, all taken from de UCI Machine Learning Repository [7].

In order to analyze the effectiveness of the OC algorithm we compared its performance against an ensemble of equal size made by C4.5. In both cases we started with the same training set. In Table 2 we summarize the results of

Task	C4.5	Our Algorithm	Reduction
balloons	0.31188	0.27938	0.10427
lymphography	0.40657	0.35122	0.13613
breast cancer	0.21871	0.04705	0.78483
wine	0.28492	0.27956	0.01881
tic-tac-toe	0.36823	0.13420	0.62023
iris	0.26000	0.23333	0.10257

Table 2. Error comparison of the traditional C4.5 and the Ordered Classification Algorithm.

the experiments for all the learning tasks. The first column presents the name of the database. Column two presents the classification error for the ensemble using the traditional learning algorithm C4.5 with a training set of 10 labeled examples. The third column shows the classification error using our algorithm when starting with 10 labeled examples and adding each iteration the 10 examples with the lowest entropy. We can see that the error was reduced in all the learning tasks when using our algorithm. The minimum reduction obtained was 1% for the learning task wine, while a maximum reduction error of 78% was reached for the learning task breast cancer. On average, for the 6 learning tasks, an error reduction of 29% was obtained using our algorithm.

In Figure 1 a graphical comparison of the experiments is presented. One point was plotted for each of the learning tasks. We plotted error rates using an ensemble of traditional C4.5 (vertical axis) against error rates using our algorithm (horizontal axis). All points lie above the diagonal line, which shows that the lowest classification errors were attained when adding the unlabeled examples to the training process.

In order to have a better understanding of the performance of our algorithm we present a detailed trace of one example. Figure 2 shows how the classification error decreases with each iteration of the OC algorithm. At the beginning of the experiment, the ensemble of C4.5 and the OC algorithm have the same classification error, as they are built with the same training set. But as the OC algorithm begins iterating the error starts decreasing. By the second iteration the error is decreased by 5% and as it keeps iterating the error continues decreasing until the last iteration, when we reach an error reduction of 90%. An ensemble of C4.5 reaches an error rate of 0.1803, compared against our algorithm, we attained an error rate of 0.0172.

6 Conclusions and Future Work

The results presented here prove that the OC algorithm can improve performance of classifiers with the use of unlabeled data. One important feature of our method is the criterion for selecting which unlabeled examples are to be used, the entropy level estimation. The entropy, as used in information theory, measures the impurity of an arbitrary

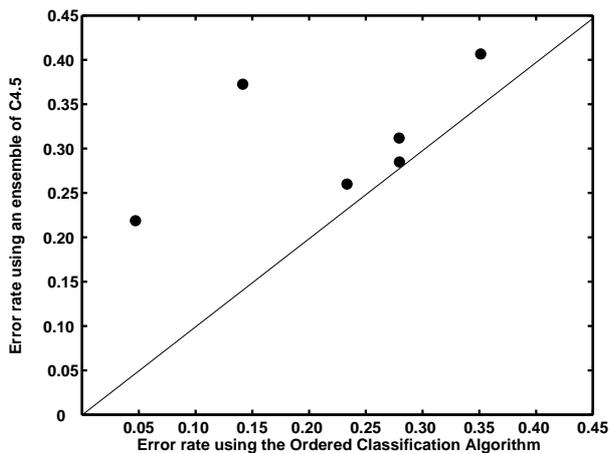


Figure 1. Comparison error between an ensemble of C4.5 and the Ordered Classification algorithm. Points above the diagonal line exhibit a lower error when using our algorithm.

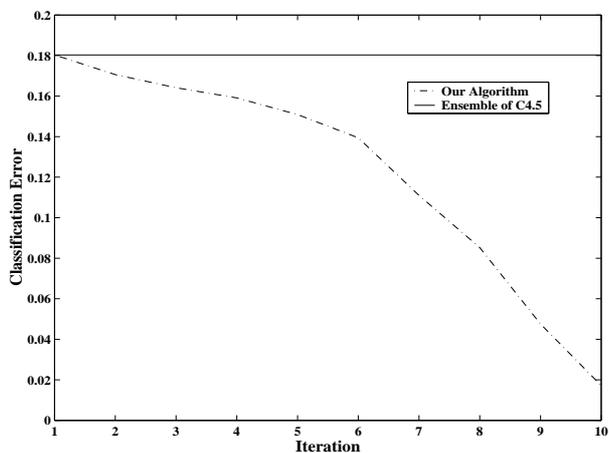


Figure 2. Error rates of the data set tic-tac-toe as they change over the course of iterations of the Ordered Classification algorithm.

set of examples. In this case, we consider the impurity level as an indication of how certain the ensemble was about the predictions it made. Even though this criterion showed to be a good approximation of the confidence level, we believe that experimenting with different methods for selecting unlabeled examples can lead us also to a reduction in the classification error.

We presented here a good alternative for building accurate classifiers. In all the experiments our algorithm outperformed the traditional supervised learning technique and error reductions of up to 78% were attained. We will continue working on new methods for improving these results. Our directions of future work include:

- Performing experiments with a different measure of the confidence level.
- Using a heterogeneous ensemble of classifiers.
- Performing experiments with other base algorithms such as neural networks.

ACKNOWLEDGEMENT: We would like to thank CONACyT for partially supporting this work under grant J31877-A.

References

- [1] F. Cozman and I. Cohen. Unlabeled data can degrade classification performance of generative classifiers. Technical Report HPL-2001-234, Hewlett-Packard Laboratories, 1501 Page Mill Road, September 2001.
- [2] K. Nigam, A. Mc Callum, S. Thrun, and T. Mitchell. Learning to classify text from labeled and unlabeled documents. *Machine Learning*, pages 1–22, 1999.
- [3] Y. Freund, H. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2/3):133–168, 1997.
- [4] S. Argamon-Engelson and I. Dagan. Committee-based sample selection for probabilistic classifiers. *Journal of Artificial Intelligence Research*, (11):335–360, Nov. 1999.
- [5] R. Liere and Prasad Tadepalli. Active learning with committees for text categorization. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pages 591–596, 1997.
- [6] J.R. Quinlan. C4.5: Programs for machine learning. 1993. San Mateo, CA: Morgan Kaufmann.
- [7] C. Merz and P. Murphy. UCI Machine Learning Repository. <http://www.ics.uci.edu/~mllearn/MLRepository.html>, 1996.

- [8] T. Solorio and O. Fuentes. Using unlabeled data to improve the automated prediction of stellar atmospheric parameters. In *Astronomical Data Analysis Software & Systems XI: ADASS XI*, 2001.
- [9] C. G. Atkeson and S. Schaal. Memory-based neural networks for robot learning. *Neurocomputing*, 9:243–269, 1995.
- [10] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. *Ninth International Conference on Information and Knowledge Management*, pages 86–93, 2000.
- [11] K. Nigam. *Using Unlabeled Data to Improve Text Classification*. PhD thesis, Carnegie Mellon University, 2001.
- [12] S. Baluja. Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data. *Neural Information Processing Sys.*, pages 854–860, 1998.
- [13] B. Shahshahani and D. Landgrebe. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. *IEEE Transactions on Geoscience and Remote Sensing*, 32(5):1087–1095, 1994.
- [14] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 1998 Conference on Computational Learning Theory*, July 1998.
- [15] S. Goldman and Y. Zhou. Enhancing supervised learning with unlabeled data. In *Proceedings on the Seventeenth International Conference on Machine Learning: ICML-2000*, 2000.
- [16] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.
- [17] R. Kohavi. Bottom-up induction of oblivious, read-once decision graphs. In *Proceedings of the European Conference on Machine Learning*, 1994.
- [18] M.T. Fardanesh and K.E. Okan. Classification accuracy improvement of neural network by using unlabeled data. *IEEE Transactions on Geoscience and Remote Sensing*, 36(3):1020–1025, 1998.
- [19] A. Blum and Shuchi Chawla. Learning from labeled and unlabeled data using graph mincuts. *ICML'01*, pages 19–26, 2001.
- [20] G. Fung and O.L. Mangasarian. Semi-supervised support vector machines for unlabeled data classification. *Optimization Methods and Software*, (15):29–44, 2001.
- [21] M. Szummer and T. Jaakkola. Kernel expansions with unlabeled examples. *NIPS 13*, 2000.
- [22] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
- [23] T. Dietterich. Ensemble methods in machine learning. In *First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science*, pages 1–15, New York: Springer Verlag, 2000. In J. Kittler and F. Roli (Ed.).