Semi-Supervised Learning of Gaussian Classifiers

Faysal Başçi

Computer Science Department

Bilkent University

Ankara, Turkey

Abstract

In this paper we present an approach that trains Gaussian classifiers using labeled and unlabeled data. Training with unlabeled data introduces efficiency in terms of time and energy spent for labeling the data. We present experiments on different data sets to illustrate the effect of unlabeled data on the performance of the classifiers. We will try to show that under specific conditions unlabeled data contains valuable information for the target function. The algorithm we utilize, first trains the classifier with limited number of unlabeled data and then with the obtained classifier labels unlabeled data and re-trains the classifier with newly labeled data and proceeds the iteration.

1. Introduction

Automatic classification of large amounts of data has become one of major daily task. For example classification of textual content is the key technology used in Internet search engines. Classification is done, mostly, by some sort of classifiers trained using data that is labeled by experts in the field. However, in most of the cases, especially when the amount of data is large, manual classification is very costly and timely. A good solution could be to utilize unlabeled data in training classifiers. In [1] Nigam et al combined expectation maximization and Naïve Bayes classifiers to classify text documents using limited amount of labeled and a large pool of unlabelled data. In their experiment they showed that an improvement of up to 30% could be obtained in classification accuracy.

Learning with both labeled and unlabeled data is known as semi-supervised learning. If the underlying probabilistic model assumption matches the data generating distribution, the reduction in variance leads to an improved classification accuracy [6]; this situation has been analyzed before by [3], [4]. However, in [6], it has been shown that when the assumed probabilistic model does not match the true data generating distribution, using unlabeled data can be detrimental to the classification accuracy; indeed this behavior observed empirically before by

[3], [1] and generally ignored or misinterpreted. This result emphasizes the importance of using correct modeling assumption when learning with unlabeled data.

In many classification problems, simple structures (e.g., the Naive-Bayes classifier [7], [8]) obtained by using just labeled data have been used successfully, such structures fail when trained with both labeled and unlabeled data [9]. Bayesian networks are probabilistic classifiers in which the joint distribution of the features and class variables is specified using a graphical model [10]. This kind of graphical representation has several advantages including the existence of algorithms for inferring the class label, the ability to intuitively represent fusion of different modalities with the graph structure [11], [12], the ability to perform classification and learning without complete data, and, most importantly, the ability to learn with both labeled and unlabeled data.

The rest of the paper is organized as follows: First we are going to discuss the value of unlabeled data, presenting results from the literature. Then, we are going to explain the method used to combine labeled and unlabeled data to train classifiers. Experimental results will follow that section and finally we are going to present the conclusion and future work.

2. Value of Unlabeled Data

Unlabeled data *alone* aren't sufficient to yield better-than-random classification in most cases because there is no information about the class label. However, unlabeled data do contain information about the joint distribution over features other than the class label. Because of this, they can be used together with a sample of labeled data to considerably increase classification accuracy in certain problems. To illustrate this, consider a simple classification problem in which instances are generated using a Gaussian mixture model. In this example, data are generated from two Gaussian distributions, one per class, whose parameters are unknown. Figure 1 illustrates the Bayes-optimal decision boundary ($x > d$), which classifies instances into the two classes shown by the shaded and unshaded areas. Note that Bayes rule provides us facilities to calculate d provided that we know the Gaussian mixture distribution parameters (*i.e.*, the mean and variance of each Gaussian, and the mixing parameter between them). Consider when an infinite amount of unlabeled data is available, along with a finite number of labeled samples. It is well known that unlabeled data alone, when generated from a mixture of two Gaussians, are sufficient to recover the original mixture components [13]. Obviously, it is impossible to assign class labels to each Gaussians without any labeled data. Consequently, the remaining learning problem is to assign class labels to the Gaussians. For example, in Figure 1, the means, variances, and mixture parameter can be learned with unlabeled data alone. Labeled data must be used to determine which Gaussian belongs to which class. This problem is known to converge exponentially quickly in the number of labeled samples [14]. In an informal manner, we can say, as long as there are enough labeled examples to determine the class of each component, the parameter estimation can be done with unlabeled data alone.

We should note that, this result depends on the critical assumption that the data indeed have been generated using the same parametric model as used in classification, something that almost certainly is untrue in real-world domains. This raises the important empirical question as to what extent unlabeled data can be useful in practice in spite of the violated assumptions.

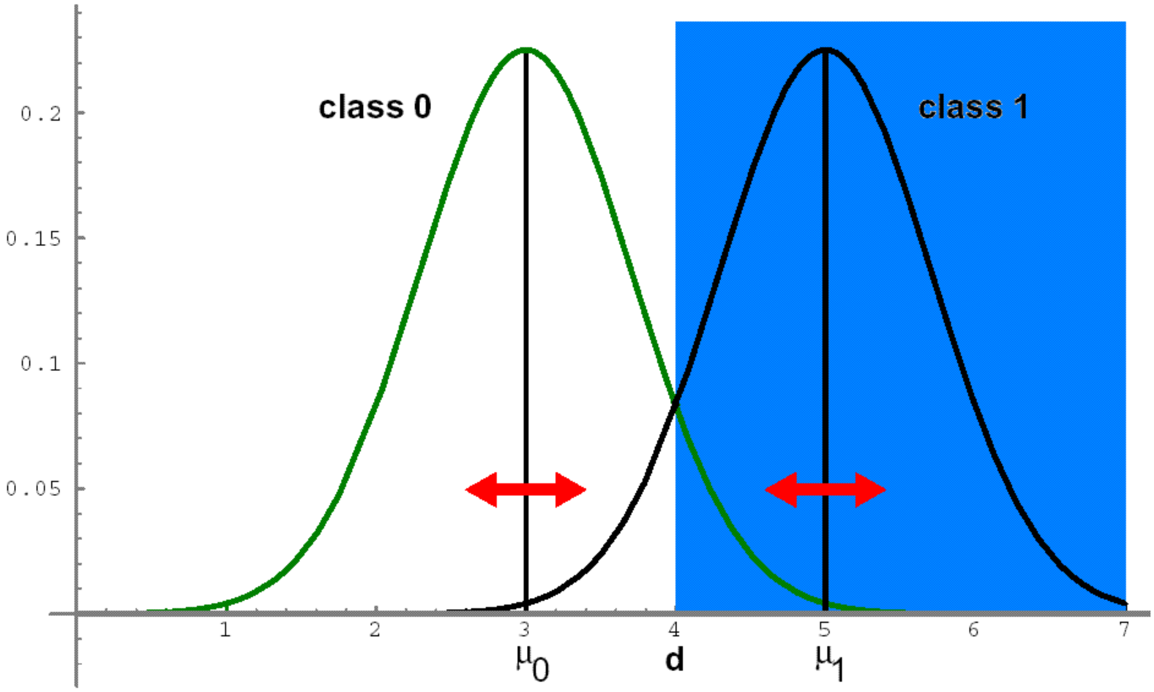


Figure 1 Classification by a mixture of Gaussians. If unlimited amounts of unlabeled data are available, the mixture components can be fully recovered, and labeled data are used to assign labels to the individual components, converging exponentially quickly to the Bayes-optimal classifier.

Now let's return to our basic problem, and present the value of unlabeled data in a more formal manner. Our goal is to classify an incoming vector of observables **X**. Each instantiation of **X** is a sample. There is a class variable **C** whose values are the classes. Our aim is to build classifiers that receive a sample **x** and output a class. Let's assume a 0-1 loss and, consequently, our objective is to minimize the probability of error (classification error). If we knew exactly the joint distribution **p(C,X)**, the optimal rule would be to choose the class value with the maximum a posteriori probability, **p(C|x)** . This classification rule achieves the minimum possible classification error, called the Bayes error. Consider the following scenario: A sample **(c, x)** is generated from **p(C,X)**. The value **c** is either known and the sample is a labeled one or the value **c** is hidden and the sample is an unlabeled one. The probability that any sample is labeled, denoted by λ, is fixed, known, and independent of the samples. Hence, the same underlying distribution **p(C, X)** generates both labeled and unlabeled data. Maximum likelihood is used to estimate $\hat{\theta}$ from a set of $\mathbf{N_l}$ labeled samples and $\mathbf{N_u}$ unlabeled samples. We consider distributions that decompose $p(C, X \mid \theta)$ as $p(X \mid C, \theta) p(C \mid \theta)$, where both $p(X \mid C, \theta)$ and $p(C \mid \theta)$ depend explicitly on θ. This is known as a generative model. The log-likelihood function of a generative model for a data set with labeled and unlabeled data is:

$$L(\theta) = L_l(\theta) + L_u(\theta) + \log\left(\lambda^{N_l}(1-\lambda)^{N_u}\right),$$

where

$$L_u(\theta) = \sum_{j=(N_l+1)}^{N_l+N_u} \log\left[p(\mathbf{x}_j|\theta)\right],$$

and

$$L_l(\theta) = \sum_{i=1}^{N_l} \log\left[\prod_C (p(C=c'|\theta)p(\mathbf{x}_i|c',\theta)^{I_{\{C=c'\}}(c_i)}\right]$$

$L_l(\theta)$ and $L_u(\theta)$ are the likelihoods of the labeled and unlabeled data, respectively. If we look carefully to above equations, statistical intuition suggests that it is reasonable to expect an average improvement in classification performance for any increase in the number of samples (labeled or unlabeled).

3. Classifier Training and Classification Method

The aim of this work is to combine a Naïve Bayes Classifier together with Expectation Maximization (EM) and observe if any performance improvement is obtained on different datasets. At time of this writing the implementation was not complete. Therefore, more detail about the detailed implementation and results will be presented in the next revision of this paper. However, in the rest of the paper the author wishes to present results obtained in similar works, and want to state that the results and analysis presented here are expected to be similar to the ones author expect to obtain after the implementation and experiments final.

The work presented here basis its theory and some key ideas to works done by [1],[2] and [6]. In [1] and [2] Nigam et al applied (EM) to Naïve Bayes Classifier. Applying EM to naive Bayes is quite straightforward. First, the naive Bayes parameters, $\hat{\theta}$, are estimated from just the labeled documents. Then, the classifier is used to assign probabilistically-weighted class labels to each unlabeled data by calculating expectations of the missing class labels, $p(C \mid X, \theta)$. Next, new classifier parameters, $\hat{\theta}$, are estimated using all the documents both the originally and newly labeled. These last two steps are iterated until $\hat{\theta}$ does not change.

4. Experimental Results

As stated in the previous section the implementation for this work is not completed yet. However, author is going to present some results from Nigam et al. The results presented are expected to be similar to the ones that is going to be obtained after the implementation and experiments are completed.

In Nigams work, they used 3 different data sets of textual content. The first one is from 20 Usenet Newsgroups. Figure 2 shows that with increasing number of unlabeled documents classification accuracy improves considerably. The second experiment is carried out on web pages of 4 University CS departments. This dataset is called WebKb dataset. The classification performance on this data set with unlabeled data is illustrated in Figure 3. The third dataset used is Reuters dataset. In this data set as documents might have more then one class a positive-negative type binary classifier is used. That is classifier decides whether a document belongs to a specific class or not. The result of the experiment is given in Table 1. It can clearly be seen in all of the experimental results that using unlabeled data increases classification performance, at least for text classification problem. For other types of problem performance improvement has also been reported as explained in the introduction part.

The aim of the author of this paper is to perform experiments on different types of datasets to observe the value of unlabeled data.

5. Conclusion

Although experiments are not complete from the results of the other authors we can say that using unlabeled data to train classifier in most of the cases improves the performance of the classifiers. This hypothesis is violated when the underlying model assumption is not correct.

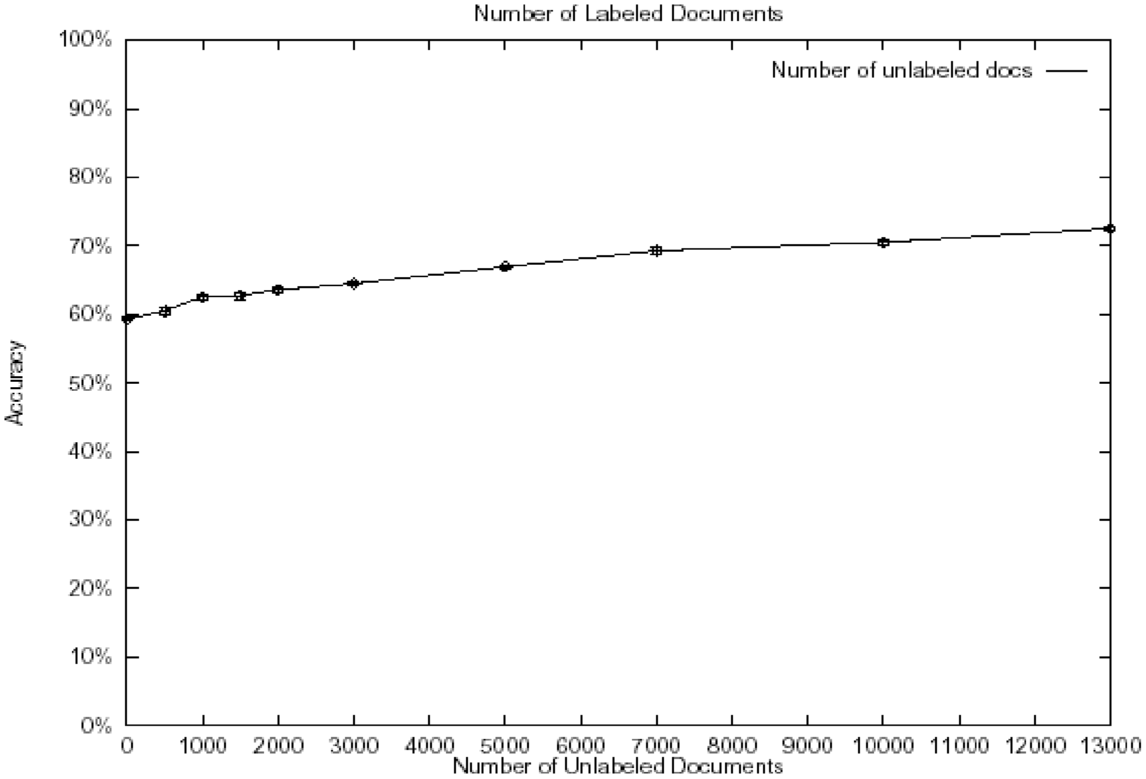


Figure 2 Accuracy of classifier with more added unlabeled data on Usenet Newsgroups. There are 400 hundred labeled documents in each case

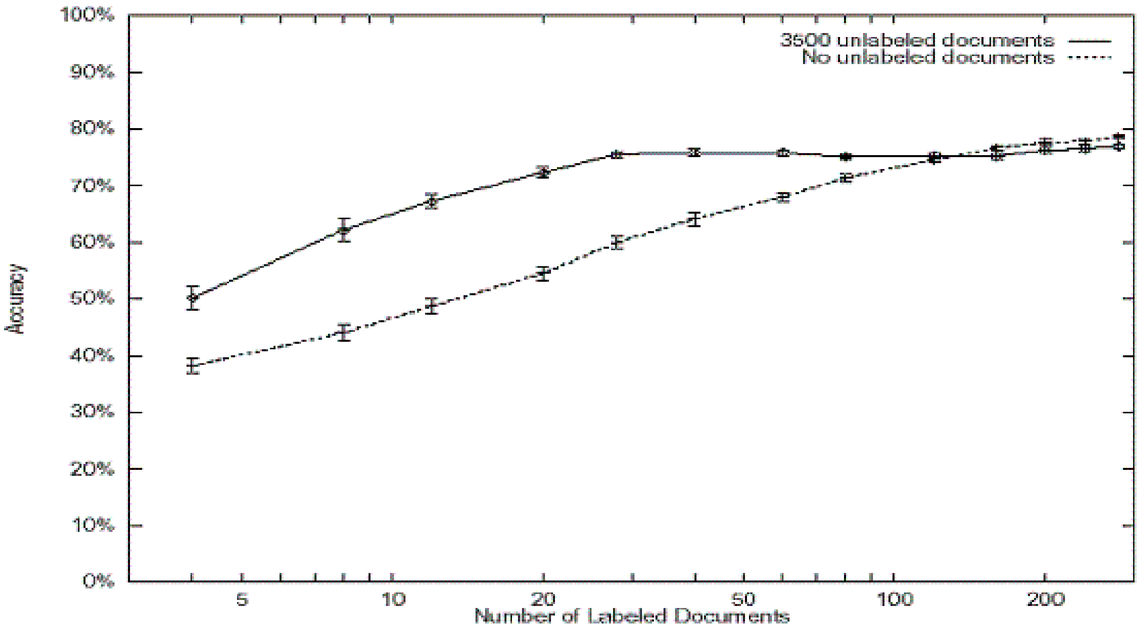
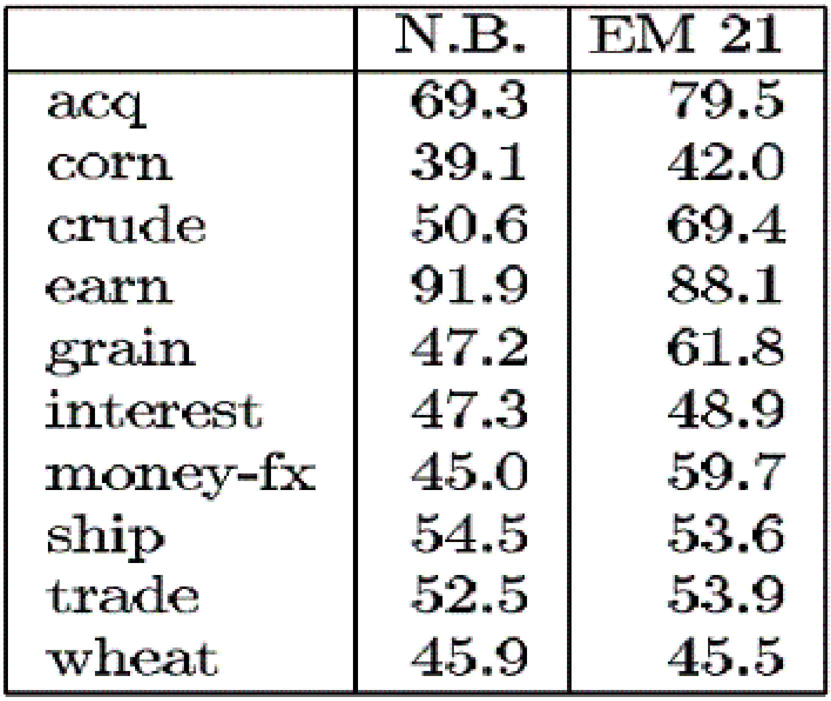


Figure 3 Comparison of accuracy of classifier with more added unlabeled data on WebKb dataset.



	N.B.	EM 21
acq	69.3	79.5
corn	39.1	42.0
crude	50.6	69.4
earn	91.9	88.1
grain	47.2	61.8
interest	47.3	48.9
money-fx	45.0	59.7
ship	54.5	53.6
trade	52.5	53.9
wheat	45.9	45.5

Table 1 Classification accuracy of classifier with fully labeled data (N.B. column) and with 40 labeled and 7000 unlabeled data (EM 21 colums)

References

[1] Nigam-2000] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. Machine Learning, 39(2/3):103-134, 2000.

[2] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Learning to classify text from labeled and unlabeled documents. In Proceedings of AAAI-98, 15th Conference of the American Association for Artificial Intelligence, pages 792-799, Madison, US, 1998.

[3] B. Shahshahani and D. Landgrebe, "Effect of Unlabeled Samples in Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon," IEEE Trans. Geoscience and Remote Sensing, vol. 32, no. 5, pp. 1087-1095, 1994.

[4] T. Zhang and F. Oles, "A Probability Analysis on the Value of Unlabeled Data for Classification Problems," Proc. Int'l Conf. Machine Learning (ICML), pp. 1191-1198, 2000.

[5] Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In Proceedings of the 11th Annual Conference on Computational Learning Theory(COLT '98), pp. 92-100.

[6] I. Cohen, F.G. Cozman, N. Sebe, M.C. Cirelo, T. S. Huang, "Semisupervised learning of classifiers: theory, algorithms, and their application to human-computer interaction," IEEE Transactions on Pattern Analysis and MAchine Intelligence, vol. 26, no. 12, pp.1553- 1566, 2004

[7] S. Baluja, "Probabilistic Modelling for Face Orientation Discrimination: Learning from Labeled and Unlabeled Data," Proc. Neural Information and Processing Systems (NIPS), pp. 854-860, 1998.

[8] R. Kohavi, "Scaling Up the Accuracy of Naive Bayes Classifiers: A Decision-Tree Hybrid," Proc. Second Int't Conf. Knowledge Discovery and Data Mining, pp. 202-207, 1996.

[9] I. Cohen, F.G. Cozman, and A. Bronstein, "On the Value of Unlabeled Data in Semi-Supervised Learning Based on Maximum-Likelihood Estimation," Technical Report HPL-2002-140, Hewlett-Packard Labs, 2002.

[10] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, Calif.: Morgan Kaufmann, 1988.

[11] A. Garg, V. Pavlovic, and J. Rehg, "Boosted Learning in Dynamic Bayesian Networks for Multimodal Speaker Detection," Proc.IEEE, vol. 91, pp. 1355-1369, Sept. 2003.

[12] N. Oliver, E. Horvitz, and A. Garg, "Hierarchical Representations for Learning and Inferring Office Activity from Multimodal Information," Proc. Int'l Conf. Multimodal Interfaces, (ICMI), 2002. pp. 219-230, 1999.

[13] McLachlan, G. J., & Krishnan, T. (1997). The EM Algorithm and Extensions. John Wiley

and Sons, Section 2.7, New York

[14] Castelli, V., & Cover, T. M. (1995). On the exponential value of labeled samples. Pattern

Recognition Letters, 16 (1), 105{111.