SEMI-SUPERVISED FISHER LINEAR DISCRIMINANT (SFLD)

Seda Remus

Clarkson University
Department of Computer Science
PO Box 5815 Potsdam, NY 13699-5815

Carlo Tomasi

Duke University
Department of Computer Science
PO Box 90129 Durham, NC 27708-0129

ABSTRACT

Supervised learning uses a training set of labeled examples to compute a classifier which is a mapping from feature vectors to class labels. The success of a learning algorithm is evaluated by its ability to generalize, *i.e.*, to extend this mapping accurately to new data that is commonly referred to as the test data. Good generalization depends crucially on the quality of the training set. Because collecting labeled data is laborious, training sets are typically small. Furthermore, it is often difficult to represent all possible observation scenarios during training, so that the statistics of the training set end up differing from those of the test data, a problem known as the sample selection bias. To address sample selection bias, we introduce a Semi-Supervised Fisher Linear Discriminant (SFLD) that utilizes additional, unlabeled data to improve generalization for both small and biased training sets. We characterize the conditions under which SFLD helps, and illustrate its benefits through experiments on digit and car recognition applications.

Index Terms— Sample Selection Bias, Generalization, Classification, Fisher Linear Discriminant

1. INTRODUCTION

The goal of a classification algorithm is to learn a mapping from feature vectors $x \in \mathcal{X}$ that describe data points to target values $y \in \mathcal{Y}$ that designate the class of the data points. The merit of any classification algorithm is evaluated based on how well it predicts the target values of previously unseen data which is measured by the generalization error, *i.e.*, the expected classification error over the underlying distribution $\mathcal{D}$ of the data. This is called the generalization problem in machine learning [3].

In this work, we investigate the quality of the labeled training data, and its effect on the generalization error. There are two main aspects of the training data quality that determine how well it can represent $\mathcal{D}$: (i) sample size, *i.e.*, number of data points in the training set; and (ii) sampling process, that is, the process for collecting or generating the training data. There is an extensive collection of past work in the literature such as [7] that studied the small sample size problem. Here, we specifically focus on the sampling process, which is the less visited aspect of training set quality. Sampling process can vary for different datasets leading to sample selection bias. More formally, a dataset has sample selection bias if it is not drawn randomly from the underlying distribution $\mathcal{D}$. Therefore, the training set may fail to represent the general population which would then cause degradation in the generalization performance.

There are many examples of semi-supervised learning in the literature that make use of unlabeled data to overcome the issues arising from the scarcity of labeled data [2], [4], [9], [5]. Most of these methods implicitly assume that the training data is an unbiased, random sample from the underlying distribution $\mathcal{D}$. There is also an active research area that specifically addresses sample selection bias. For instance, in statistics sample selection bias has been studied under the missing data problem [11], [12], which categorizes the type of the bias based on the mechanism that causes the missing data. This bias categorization has also been adopted in machine learning with particular attention on the type of bias that is caused by the features. Most of the related work in this area propose using techniques that are essentially based on either importance sampling [15] or expectation-maximization (EM) [14] methods.

It is not straightforward to determine the presence of bias in a real dataset without knowing the actual experimental setup or the data collection process thoroughly. Most real datasets are complex enough that the distinctions between different conditions under which datasets are generated may become hard to characterize. Therefore, a successful semi-supervised algorithm should not depend on assumptions relating to the sampling process. Ideally, a successful semi-supervised algorithm should be able to maintain the level of performance of a supervised learner when the labeled samples are plentiful and representative; improve performance when the labeled samples are few yet representative; and also improve performance when the labeled samples are biased, without human supervision or modification of tuning parameters. In accordance with the above description, we propose a semi-supervised Fisher linear discriminant (SFLD) unlike the existing methods as it does not require distinguishing the type of the bias or even the presence thereof. Therefore, SFLD attempts to incorporate the information from the available unlabeled samples to the learning process to improve classification accuracy when the labeled data is sparse or is biased. Section 2 describes SFLD criterion and the procedure for optimization using unlabeled data. Section 3 provides experimental results from the digit and car datasets to illustrate superior performance of SFLD compared to other semi-supervised methods. In Section 4, we conclude by discussing SFLD and potential future improvements to it.

2. OPTIMIZING SFLD USING UNLABELED DATA

We propose a new optimization criterion based on extending the original Fisher criterion to also incorporate the unlabeled data points for learning a classifier. In the context of binary classification where $\mathcal{C} = \{c_0, c_1\}$ refers to the two distinct categories, it is known that Fisher Linear Discriminant (FLD) [6] finds a weight vector w that optimizes the class separation criterion $J(w) = \frac{w^T S_B w}{w^T S_W w}$ in the labeled data where $S_B = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T$ is the between class covariance matrix and $S_W = \sum_{x \in c_1}(x - \mu_1)(x - \mu_1)^T + \sum_{x \in c_0}(x - \mu_0)(x - \mu_0)^T$ is the within class covariance matrix. Our goal is to extend this criterion so that a potential class separation in the unlabeled data can also be recovered.

The intuition behind SFLD is to move the original FLD line w in a direction such that the new decision boundary passes through a gap within the unlabeled points while keeping the separation in the labeled data points high, albeit not maximum as in FLD. Here, w is initialized to maximize the original Fisher criterion $J_L(w) = \frac{w^T S_{L_B} w}{w^T S_{L_W} w}$ on the labeled data where the subscript L stands for the labeled dataset. The new SFLD criterion $J(w) = \lambda J_L(w) + (1 - \lambda) J_U(w)$ is a convex combination of $J_L(w)$ and $J_U(w)$ where the term $J_U(w)$ serves for exploring a separation in the unlabeled set and λ is the tradeoff parameter that controls the degree to which $J_U(w)$ is optimized versus $J_L(w)$. Moreover, we define $J_U(w)$ to be:

$$J_U(w) = \frac{w^T \hat{S}_{U_B} w}{w^T \hat{S}_{U_W} w} = \frac{(\hat{\mu}_{-U} - \hat{\mu}_{+U})^2}{(\hat{\sigma}^2_{-U} + \hat{\sigma}^2_{+U})} \tag{1}$$

where $\hat{\mu}$ and $\hat{\sigma}$ denote the maximum likelihood estimates of the class conditional parameters and $\hat{S}_{U_B}$ and $\hat{S}_{U_W}$ are the between and within class scatters for the unlabeled data. These estimates are found using the unlabeled dataset $\{(x_{U_k})\}_{k=1,\dots,N_U}$ and the predicted labels $\{(\hat{y}_{U_k})\}_{k=1,\dots,N_U}$ inferred by w. To maximize $J(w)$, we propose a gradient ascent based iterative optimization method. Since the label estimates $\{(\hat{y}_{U_k})\}_{k=1,\dots,N_U}$ are based on the current w, each iteration has the following two steps: (1) Predict the labels according to the current w and update $\hat{\mu}$ and $\hat{\sigma}$ of each class according to the current predicted labels; (2) Update w in the direction of the gradient of $J(w)$.

Since this optimization scheme is based on shifting the original w until a gap in the unlabeled data is found, w is initialized to $w^{(0)}$, the original vector that maximizes the Fisher criterion $J_L(w)$ for the labeled training data. At each iteration i, first the labels of the unlabeled data are estimated using $w^{(i-1)}$, and these estimated labels $\{(\hat{y}_{U_k})\}^{(i)}_{k=1,\dots,N_U}$ are used to update the class conditional parameters $\hat{\mu}^{(i)}$ and $\hat{\sigma}^{(i)}$ of the unlabeled data. For an unlabeled feature vector x_U, the label $\hat{y}^{(i)}_U$ assigned at iteration i is estimated as:

$$\hat{y}^{(i)}_U = \begin{cases} 1 & \text{if } {w^{(i-1)}}^T x_U \geq 0, \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Using these predicted labels, the maximum likelihood estimates $\hat{\mu}^{(i)}$ and $\hat{\sigma}^{(i)}$ for the class conditional parameters of the positive and negative classes of the unlabeled data can be computed.

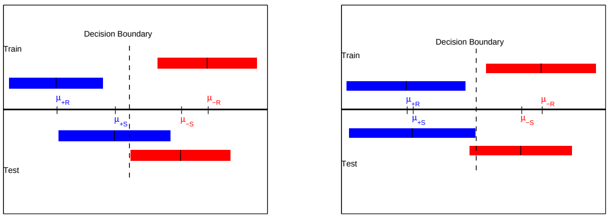
In the second step of iteration i, the new vector $w^{(i)}$ is computed by updating $w^{(i-1)}$ in the gradient direction $g^{(i-1)}$. Thus, $w^{(i)} = w^{(i-1)} + \alpha^{(i)} g^{(i-1)}$. Here, the optimal step length $\alpha^{(i)}$ is determined automatically. We set $\alpha^{(i)}$ to the α that maximizes $J(w^{(i)}) = J(w^{(i-1)} + \alpha^{(i)} g^{(i-1)})$ when $w^{(i-1)}$ and $g^{(i-1)}$ are held constant. This maximization is performed by finding the roots of the quadratic polynomial that maximizes the above expression.

Another parameter that has a role in maximizing $J(w)$ is λ, the tradeoff parameter that balances the effect of the labeled and unlabeled datasets on the optimization function. This parameter can have a significant influence on the step size α. Therefore, instead of setting λ arbitrarily, we decided that $\frac{J_U(w^{(0)})}{J_L(w^{(0)}) + J_U(w^{(0)})}$ can be a simple and useful measure for balancing the effects of the labeled and unlabeled data. Finally, when the difference between $w^{(i)}$ and $w^{(i-1)}$ falls below a certain threshold, w is no longer updated.

3. RESULTS

To illustrate how classification is affected by using training and test sets that are collected under different conditions, we will use the car detection example with datasets obtained from LabelMe [13] and UIUC [1] databases. These two databases are similar in the sense that they both have images that contain a side view of a car cropped inside a rectangular bounding box. On the other hand, one obvious difference between the two databases is the scale and resolution of the car images.

Figure 1 shows two cases in which both test sets are constructed using the UIUC database, with different training set formations. The diagram on the left refers to the case where the training set is formed using only LabelMe; whereas the diagram on the right refers to the case where the training set is formed using a combination of LabelMe and UIUC databases. The reference line in the diagram is used to indicate the line found by FLD whose direction vector is denoted as w. This line is used for projecting the data onto a one dimensional subspace while maximizing the separation criterion $J(w)$ of FLD. In each diagram, the top of the reference line corresponds to the training set whereas the bottom one corresponds to the test set. The colored bars represent two standard deviations of the projected data centered symmetrically around the mean of each of the classes. The blue bars correspond to the positive (car) class, and the red bars correspond to the negative (non-car) class. The training class-conditional means are denoted by μ_{+R} and μ_{-R}, whereas the test means are denoted by μ_{+S} and μ_{-S} for the positive and negative classes respectively as the subscripts $+$ and $-$ refer to the classes and R and S refer to the training and test sets. The dashed line indicates the decision boundary that minimizes classification error on the training set. It is clear from Figure 1 (a) that when training and test datasets are obtained from different image sources such as LabelMe and UIUC, the subspace with maximum separation for the training set is far from separating the classes for the test set. This is visible from the significantly overlapping red and blue bars in Figure 1 (a). By contrast, the overlap between the two classes for the test set is significantly reduced in Figure 1 (b) compared to the first case since there are also images from the UIUC database in this training set.



(a) (b)

Fig. 1. Projection of car data when FLD is trained using (a) only the LabelMe dataset; (b) the union of LabelMe and UIUC datasets. The bars above the reference line show the training set and the ones below show the test set (UIUC) projections.

We also verified that SFLD is able to significantly improve classification accuracy when the training and test sets are not similar. Here, car and digit recognition applications are used as examples for comparing the performance of SFLD with some other prominent works in this area.

(a) LabelMe example (b) UIUC example

Fig. 2. Examples from car databases

(a) MNIST Digit 4 (b) USPS Digit 4

Fig. 3. Examples from digit databases

For the car recognition example, we used the same image databases that we used for the diagrams in Figure 1. Therefore, the LabelMe database is used as the labeled training set, and the UIUC database is used as the unlabeled data for forming the test set. Examples of car images from these databases are shown in Figure 2. For the digit recognition example, the task is to distinguish the digit 4 from the digit 9 since we focus on binary classification in this work. We picked the digits 4 and 9 as they are somewhat similar in shape, thus it is relatively more challenging to classify them. The datasets that we use here are selected from the National Institute of Standards and Technology (MNIST) [10] and the US Postal (USPS) [8] digit databases, which are used as the labeled training set and the unlabeled set respectively. Figure 3 shows several digit images from these databases.

Table 1 summarizes the classification errors for the car and digit recognition applications when the test set is a subset of the unlabeled data, and the rest of the unlabeled data is used for optimizing SFLD. Thus, the unlabeled data that is used for optimization is different than the unlabeled data that is used for testing. Here, we compare our method to the original Fisher linear discriminant (FLD), transductive Support Vector Machine (tSVM) [9], and the weighted Fisher linear discriminant (WFLD) where tSVM and WFLD methods also use unlabeled data for improving classification accuracy. The weights in WFLD correspond to the density ratios $p(x|s = 0)/p(x|s = 1)$ which is a standard weight used in the sample selection bias literature [15].

These error rates are obtained from the average of five random partitions of the unlabeled data, where 30% of the unlabeled data is used for testing. It is clear that for both digit and car recognition, there is significant improvement in classification performance when SFLD is used for utilizing the information from the unlabeled data. Moreover, both tSVM and WFLD fail to improve the classification error on these recognition tasks where the labeled training data and the unlabeled data are substantially different. Semi-supervised techniques that are based on reweighing, such as WFLD, are only viable when the labeled and unlabeled data overlap significantly. Otherwise, the density ratios $p(x|s = 0)/p(x|s = 1)$ can not be estimated reliably. Therefore, the dissimilarity between the labeled and unlabeled datasets that are used here can account for the failure of WFLD. On the other hand, tSVM takes as input the ratio between the number of points in

Method	Test Dataset	Error Rate
SFLD	USPS Digit	10.48 %
FLD	USPS Digit	20.87 %
tSVM	USPS Digit	20.42 %
WFLD	USPS Digit	32.62 %
SFLD	UIUC Car	11.40 %
FLD	UIUC Car	18.29 %
tSVM, $r = 0.15$	UIUC Car	26.09 %
tSVM, $r = 0.5$	UIUC Car	11.75 %
WFLD	UIUC Car	23.35 %

Table 1. Classification accuracies when the test set is a hold-out sample from the unlabeled dataset

two classes, in order to avoid an unbalanced split. This is one of the undesirable properties of the tSVM scheme as it is often not realistic to know this ratio a priori. The convention is to use the same ratio obtained from the training set, however this can be misleading especially when there is sample selection bias. According to our experiments, this parameter can have a significant impact on performance. In the digit recognition example, when the more accurate class ratio $r = 0.5$ is given, tSVM and our method performs similarly; whereas when $r = 0.15$, there is significant drop in the performance of tSVM.

4. DISCUSSION

This work investigated the effect of using training and test sets that were collected under different conditions on the classification performance. In this context, car and digit recognition tasks were used to illustrate how the information from the labeled training data can be insufficient for locating the decision boundary in the test data. In order to simulate different conditions for data collection, we used training datasets that were obtained from a different image database than the test datasets.

We proposed the semi-supervised Fisher linear discriminant (SFLD) to utilize the information from the relatively more abundant unlabeled data in order to bridge the gap between the insufficient labeled data and the test data. In digit and car recognition, SFLD successfully improved classification accuracies by locating the region in the test data where there is more prominent class separation. However, it should be noted that there are also several assumptions for SFLD to succeed. One of the main restrictions is the requirement of linear separability as FLD is the classifier that we build on. Therefore, extending a similar idea to non-linear classifiers would be an interesting study in the future. Kernel methods can provide insight here by making data linearly separable in higher dimensions, possibly evoking more challenges at the same time. Also, since SFLD is an iterative method, it can possibly converge to a local maximum. Transforming the feature space to eliminate regions where within class gaps exist can help alleviate this problem.

References

[1] S. Agarwal, A. Awan, and D. Roth. The UIUC image database for car detection. `http://l2r.cs.uiuc.edu/∼cogcomp/Data/Car/`, 2002.

[2] M. Belkin and P. Niyogi. Using manifold stucture for partially labeled classification. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 929–936. MIT Press, 2003.

[3] C. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.

[4] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, 2006.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[6] R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[7] K. Fukunaga and R. R. Hayes. Effects of sample size in classifier design. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(8):873–885, 1989.

[8] J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994.

[9] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML '99: Proceedings of the Sixteenth International Conference on Machine Learning*, pages 200–209, 1999.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998.

[11] R. J. A. Little and D. B. Rubin. *Statistical analysis with missing data*. New York: John Wiley, 2002.

[12] D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[13] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: a database and web-based tool for image annotation. *MIT AI Lab Memo AIM-2005-025*, 2005.

[14] A. Smith and C. Elkan. Making generative classifiers robust to selection bias. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 657–666, 2007.

[15] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, pages 114–221, 2004.