Semi-supervised Object Detector Learning from Minimal Labels

Sudeep Pillai
spillai@csail.mit.edu

Abstract

While traditional machine learning approaches to classification involve a substantial training phase with significant number of training examples, in a semi-supervised setting, the focus is on learning the trends in the data from a limited training set and simultaneously using the trends learned to label unlabeled data. This report focuses on a particular semi-supervised learning technique called Graph-based Regularized Least Squares (LapRLS) that can learn from both labeled and unlabeled data to build a robust object detector capable of learning purely from a single training example and a reasonably large set of unlabeled examples. The algorithm employs multiple feature descriptors such as GIST and Pyramid Histogram of Oriented Gradients (PHOG) for learning these trends. Performance comparisons against traditional supervised learning algorithms demonstrate that LapRLS outperforms the supervised classifiers on several datasets especially when the number of training examples are minimal. As a particular application, the LapRLS performs considerably well on Caltech-101, an object recognition dataset.

1. Introduction

In a setting where labeled data is hard to find or expensive to attain, we can formulate the notion of learning from the vast amounts of unlabeled instances in the data given a few minimal labels per class instance. Formally, semi-supervised learning addresses this problem by using a large amount of unlabeled data along with labeled data to make better predictions of the class of the unlabeled data. Fundamentally, the goal of semi-supervised classification is to train a classifier \( f \) from both the labeled and unlabeled data, such that it is better than the supervised classifier trained on the original labeled data alone. Several object learning techniques [2], [7] based on semi-supervised frameworks have shown convincing results for classification with minimal labels.

Semi-supervised learning has tremendous practical value in several domains [11] including speech recognition, protein 3D structure prediction, video surveillance etc. In this report, the primary focus is to learn trends from labeled image data that may be readily available from human annotation, or some external source, and using the learned knowledge to label the evergrowing data deluge of unlabeled images on the internet. Particularly, the focus is on utilizing these semi-supervised techniques on Caltech-101 [4], [5], an object recognition dataset while also providing convincing results on toy datasets.

2. Background and related work

Before we delve into the details of the motivation and implementation behind semi-supervised learning, it is important to differentiate two distinct forms of semi-supervised learning settings. In semi-supervised classification, the training dataset contains some unlabeled data, unlike in the supervised setting. Therefore, there are two distinct goals; one is to predict the labels on future test data, and the other goal is to predict the labels on the unlabeled instances in the training dataset. The former is called inductive semi-supervised learning and the latter transductive learning [12].

While it is reasonable that semi-supervised learning can use additional unlabeled data to learn a better predictor \( f \), the key lies in the model assumptions about the relation between the marginal distribution \( P(x) \) and the conditional distribution \( P(y|x) \).

![Figure 1. (a) Plots visualizing the cluster assumption, (b) and the cluster/manifold assumption](image)

In this semi-supervised scenario, the idea is to use a regularizer that prefers functions which vary smoothly along the manifold and do not vary in high density regions as
from graph theory, the combinatorial Graph Laplacian, is the matrix representation of a graph. The Graph Laplacian is defined as

$$L = D - W$$  (1)

where \( D_{ii} = \sum_j W_{ij} \) is the degree of node \( i \) and \( W \) is the weight matrix as defined above. An important point to note is that \( f \) is constrained to take values \( f(i) = y_i, i \in L \) on the labeled data. Using the laplacian, one can define the energy function for a graph by the following:

$$E(f) = \frac{1}{2} \sum_{i,j} w_{ij} (f(i) - f(j))^2 = f^T L f$$  (2)

Minimization of the smoothness term \( f^T L f \) can be achieved by the trivial solution of \( f = 1 \), but in a semi-supervised setting such as this one, the minimization is a combination of the smoothness and the training loss(label agreement). For the squared training loss, this is defined as

$$J(f) = \frac{f^T L f + (f - y)^T \Lambda (f - y)}{\text{smoothness} \quad \text{label agreement}}$$  (3)

where \( \Lambda \) is the diagonal matrix whose diagonal elements are \( \Lambda_{ii} = 0 \) for \( i \in U \) (unlabeled points), and \( \Lambda_{ii} = \lambda \) for \( i \in L \) (labeled points). The final solution has a closed form to the following equation \((L + \Lambda)f = \Lambda y\) given by:

$$f = (L + \Lambda)^{-1} \Lambda y$$  (4)

Although the solution can be given in closed form for the squared error loss, it requires a solution to an \( n \times n \) linear system which poses serious problems when \( n \) grows larger. While this report doesn’t focus on larger datasets, it is important to note that various approximations to the final solution exists for larger \( n \), specifically as suggested in [6].

4. Experiments and Results

In the following section, the performance of the Graph-based Laplacian Regularized Least Squares (LapRLS) algorithm is evaluated and tested on a variety of toy datasets and Caltech-101 [4], [5], an object recognition dataset.

4.1. Toy Datasets: 2 Moons Dataset

In order to guage and visualize the performance of these aforementioned algorithms, a few toy datasets were generated. Since the data is represented in 2D, we use the data as is and build the weight matrix, and laplacian matrix for the data. The edge weights are determined using an RBF with \( \sigma = 0.35 \) for the case of the SVM-RBF Kernel, and the LapRLS.

A classical dataset for classification is the 2 moons dataset, where the datapoints can be easily classified by visualizing the data, but non separable with linear kernel.
SVMs, and RLS. The plots in figure 3 below show the performance of each of the classifiers discussed, with LapRLS outperforming especially when the number of training examples is minimal.

4.1.1 Performance compared to \(k\)NN classifiers

It is of particular interest to understand why the performance of the LapRLS with an RBF kernel is significant compared to the other classifiers in this case. Since there is continuity in the data manifold, one can propagate labels accordingly by minimizing the overall objective function keeping the regularity in manifold consistent. As compared to a \(k\)NN classifier, the LapRLS avoids labeling the second manifold (data from a different class) with the same label as itself since there are high derivatives near regions where manifolds come close to each other. Figure 4 compares a traditional \(k\)NN classifier with the graph-based LapRLS for different values of \(k\). The plots clearly show that the RBF-LapRLS outperform the \(k\)NN classifier in most cases (both in cases where \(N_t\) is low and high). In the case of \(k\)NN classifiers, the high misclassification error rate can be attributed to outliers in the data that tend to propagate labels undesirably, leading to overall misclassification of the data.

4.1.2 Performance with Minimal Labels

When only a few training examples are available, both linear-SVMs and linear-RLS tend to find a half-plane that best classifies the training examples without any notion of utilizing full information. The linear-SVM classifier was modified to include a slack variable for datasets that were linearly non-separable. Both these classifiers show similar results, but overall poor performance on the 2 moons dataset. While RLS and linear-SVM have relatively low misclassification error with fewer training examples \(N_t\) in this case, as \(N_t\) increases, the misclassification rates tend to increase.

In the case of the RBF-SVM (3rd column in figure 4), performance tends to increase with increasing \(N_t\). This is expected as the classifier tends to overfit to the training data, and as \(N_t\) approaches \(N\) (full test/train dataset size), the performance should converge to optimal. As a side note, in order to compare the performance between the RBF-SVM and the RBF-LapRLS, both variations were provided with the same kernel parameters. The interesting performance measures are when the training size \(N_t\) is minimal (1st and 2nd row). Here we see that the LapRLS outperforms the RBF-SVM by correctly classifying all the unlabeled points in the dataset, while the RBF-SVM uses only the training examples to make classification decisions. This results in poor performance and lack of generalization for the specific dataset as seen in rows 1, 2, and 3 in figure 4 (columns 3 and 4 comparing RBF-SVM and LapRLS). The LapRLS performs substantially better than any of the classifiers compared, even with minimal data, or with full training data \(N_t\).

In short, utilizing unlabeled data does not necessarily hurt the classification as long as the data belongs to either of the classes in question. If however, spurious data points that correspond to a class that is not well represented or labeled in the training set tends to bias the classification, especially when their manifolds overlap or approach each other.
4.2. Caltech-101

In order to test the true performance and utility of the Graph-based Laplacian Regularized Least Squares (LapRLS) approach, the algorithm was tested on Caltech-101, an object recognition dataset. The dataset consists of 101 classes, each of which contains approximately 800 images per object class. Since we limit ourselves to understanding the trends in the data using minimal labels, only 7 classes (Faces, Airplanes, Motorbikes, Cars, Sunflower, Chair, Emu) were considered in our experiments and tests.

4.2.1 Feature extraction and classification

To capture the main essence of the object in the image, both the GIST descriptor [8] and the Pyramid Histogram of Oriented Gradients (PHOG) descriptor [1] was used separately to benchmark their capabilities. GIST, a 512-dimensional descriptor has a lot of mutual information between descriptors when extracted for different images. To this end, its dimensionality is reduced to 64 dimensions via PCA, while still retaining over 98% of the variance in the data. PHOG, on the other hand is extracted by building a vocabulary of words/histograms (600 words in our case), to characterize the objects in the image.

As expected, each features extracted via GIST and PHOG independently lie on a feature space that is specific to the feature extraction technique. As explained earlier, semi-supervised learning makes an assumption that these features extracted from images need to lie on a continuous manifold. [3], [8] suggests that euclidean distance is a sufficient metric to compute the similarity measure between GIST descriptors of images. [1] suggests that $\chi^2$ distance as a good metric to measure the similarity between PHOG descriptors of images. While other distance metrics may work equally as well, the results look promising regardless of the distance metric that was used.

A one-versus-all approach was taken to classify the 7 classes from each other. Here each classifier $f_k$, solution to eqn. 4, would be responsible for predicting those examples that fall within the $k$th class as positive and classifying the rest of the dataset as negative. In such a scenario, the one-versus-all formulation is given by $f(x) = \arg \max_i f_i(x)$, breaking ties arbitrarily.

4.2.2 Graph-based Laplacian RLS

Once the features are extracted and reduced in their dimensionality, the similarity measure is determined between each pair of nodes in the dataset to build an $n \times n$ weight matrix as discussed in section 3. The LapRLS again takes an RBF-kernel to account for non-linear but continuous manifolds on which these feature descriptors lie on. Figures 6, 7 show the performance of LapRLS (with GIST and PHOG independently) versus other kNN classifier ($k = 2, 3, 6$) with increasing number of training examples $N_t$. Once again, LapRLS outperforms each of the kNN classifiers for most of the cases for $N_t = 20$ (Average per-class misclassification error $\sim 10\%$ and $\sim 13\%$ for GIST and PHOG respectively), except for $N_t = 1$ (Average per-class misclassification error $\sim 43\%$ and $\sim 27\%$ for GIST and PHOG respectively). This may be attributed to the fact that the class-specific manifolds each of these GIST and PHOG features lie on may not be entirely continous (or have high derivatives) causing the regularization of the label assignment for the LapRLS to have poor performance. It could also be the case that the specific training example that was provided for the learning was not close to the edge of a manifold, making the discrimination between classes difficult as the labels are propagated farther away from the original training set.

4.2.3 Performance with Minimal labels

The following figures 6, 7 show the performance of the LapRLS algorithm with increasing number of training examples using GIST and PHOG descriptors. The figure to the left shows the Receiver-Operating-Characteristic (ROC) curve for the performance of the object recognition task. The curves tend to imply that with increasing number of training examples, $N_t$, the performance continues to improve. More importantly, the LapRLS has considerable performance even with $N_t = 1$ (blue curve), and continues to rise when $N_t = 40$ (orange curve). The figure on the right shows the confusion matrix for the object recognition task.

The following figure (Figure 8) shows the results of the application of the Graph-based Laplacian Regularized Least Squares framework with GIST features in an object recognition setting.

The LapRLS works as expected with promising performance even with limited training data samples. It is of particular interest to realize the regions of poor performance for the LapRLS case. In the case of faces or motorbikes in the figure 8, the faces and motorbikes that could not be
Figure 6. Plot showing the performance of LapRLS on Caltech-101 using GIST descriptors. (a). Comparison against kNN classifiers with increasing number of training examples. Once again, LapRLS seem to perform better for both small and larger \( N \) training samples. (b). ROC curve showing the performance of LapRLS on Caltech-101 with increasing number of training examples \( N \). The figure shows that LapRLS has considerable performance even with 5 training examples. (c). Confusion matrix of the multi-class classification on Caltech-101 showing reasonably good performance on training with a global descriptor such as GIST, and a single training example.

Figure 7. Plot showing the performance of LapRLS on Caltech-101 using PHOW descriptors. (a). Comparison against kNN classifiers with increasing number of training examples. Once again, LapRLS seem to perform better for both small and larger \( N \) training samples. (b). ROC curve showing the performance of LapRLS on Caltech-101 with increasing number of training examples \( N \). The figure shows that LapRLS has considerable performance even with 5 training examples. (c). Confusion matrix of the multi-class classification on Caltech-101 showing reasonably good performance on training with a global descriptor such as GIST, and a single training example.

classified correctly (column 4) seem to have a significantly different background than the training set (column 1) or the images that have been classified correctly (column 2). This may be due to the fact the feature descriptor that is used to discriminate between images captures the background information as well (or possibly looks at contrasting foreground/background scenes). That said, it would make sense that the objects in the 4th column do not correlate too well with the original training example where the background seems to be less cluttered. This may be accounted for in the training process where more optimal training labels can be actively suggested to the user for annotation while simultaneously optimizing the overall classification objective function. This sub-domain is referred to as active learning [9] and has received quite a bit of reception from the computer vision community in the recent years.

Of particular interest is when one can combine the capabilities of multiple feature spaces to bootstrap an overall better object classification system. In our case, while GIST and PHOG were independently benchmarked, its performance together with a unified kernel [10] can prove to be highly beneficial when one of the descriptors fail to classify correctly in regions where the other descriptor performs well. While the kernelization of multiple descriptors were attempted, much convincing results could not be produced before the conclusion of this report.

5. Conclusion

To summarize, it is demonstrated that a semi-supervised framework can be very beneficial specifically in object
recognition, especially when the number of training data is limited. Furthermore, from the results it is evident that the aforementioned graph-based laplacian least squares can leverage the availability of unlabeled data in building a better classifier as compared to the traditional supervised classifiers such as RLS, Linear-SVM and RBF-SVM. While this report predominantly focuses on the performance of such classifiers when the training data is limited, the results also generalize quite well with increasing number of training examples and still continue to be the lower bound of misclassification error rate from figures 4, 6, and, 7. As a final note, this report incorporates a novel Graph-based Regularized Least Squares semi-supervised learning technique for multiclass-object recognition using GIST and PHOG features, providing convincing recognition performance despite a minimal set of training examples.

References


