Semi-Supervised Learning of Speech Sounds

Aren Jansen, Partha Niyogi

Department of Computer Science, University of Chicago, Chicago, IL USA
aren@cs.uchicago.edu, niyogi@cs.uchicago.edu

Abstract

Recently, there has been much interest in both semi-supervised and manifold learning algorithms, though their applicability has not been explored for all domains. This paper has two goals: (i) to demonstrate semi-supervised approaches based solely on clustering are insufficient for phoneme classification and (ii) to present a new manifold-based semi-supervised algorithm to remedy this shortcoming. The improved performance of our approach over cluster-based methods substantiates the practical relevance of a geometric perspective on speech sounds.

Index Terms: semi-supervised learning, speech sound classification, manifold learning algorithms

1. Introduction

The goal of semi-supervised learning is to incorporate unlabelled data to improve classification accuracy relative to equivalently complex fully-supervised methods when both are provided equal numbers of labelled training examples. This has particular significance in speech recognition applications where labelled data may be expensive to obtain (e.g. novel languages, speakers, and tasks). Underlying each semi-supervised learning algorithm is an implicit assumption about the structural nature of the input data classes. A clustering assumption is a straightforward choice, employed by the semi-supervised expectation-maximization for Gaussian mixture models (EMGMM) algorithm [1]. However, the clustering model breaks down in applications where the data classes exhibit more complex spatial relationships. Furthermore, the introduction of unlabelled data under a faulty structural assumption can degrade classification performance relative to the original fully-supervised method.

In the case of certain classes of speech sounds, whose short time Fourier representations have been shown to lie on or near a non-Euclidean manifold [2][3], we propose that a more general structural assumption must be adopted. The only existing coherent manifold-based framework for semi-supervised learning is manifold regularization [4], based on the theory of the graph Laplacian operator. However, this approach involves extensive model selection that necessitates large labelled validation sets. In the case of phoneme classification, where dataset annotation is particularly costly, a manifold-based semi-supervised approach that requires minimal supervision (e.g., 2-20 examples) is more desirable.

This minimal supervision regime requires an algorithm with either no model selection or a method for automatic parameter selection. Toward this end, we present a new approach, kernelized locality preserving projection-based semi-supervised learning (LPPSSL), which also relies on the power of the graph Laplacian operator to estimate an intrinsic basis for the manifold. Our goal in this paper is to demonstrate that LPPSSL's ability to accommodate both manifold and cluster structure is essential to successful incorporation of unlabelled data. To accomplish this, we present a series of carefully constructed toy classification experiments using both LPPSSL and EMGMM that isolate the utility of our geometric perspective. These experiments are a first step in clarifying the fundamental issues involved in semi-supervised learning in speech. We believe that such fundamental understanding is necessary for the successful incorporation of semi-supervised methods in practical systems.

2. The LPPSSL Algorithm

The LPPSSL algorithm involves three steps: (1) construct the graph Laplacian matrix using unlabelled data, (2) determine an intrinsic basis for the manifold using a modified version of kernelized LPP [5], and (3) determine a map from that intrinsic basis to the class labels using a small set of labelled training examples.

2.1. The Graph Laplacian Operator

We assume speech data lies on a Riemannian manifold $\mathcal{M}$ for which the Laplacian, $\Delta_{\mathcal{M}}$, is the second order differential operator. It is a positive semidefinite operator whose eigenfunctions form an orthogonal basis for $\mathcal{L}^2(\mathcal{M})$ [6]. An eigenfunction's smoothness on the manifold is determined by the magnitude of the corresponding eigenvector. Therefore, if we limit an eigenbasis expansion of a function to finite terms, we can impose any desired level of smoothness in the approximation. Furthermore, each eigenfunction varies smoothly with *geodesic* distance on $\mathcal{M}$ and is therefore faithful to the geometry of the manifold.

In practice, we are not given an analytical form of the manifold $\mathcal{M}$, so the Laplacian operator cannot be used directly. Instead, we must implement the graph theory analogue as follows: Consider a manifold $\mathcal{M}$ embedded in $\mathbb{R}^H$ and N data points $x_1, \ldots x_N \in \mathcal{M}$. We can construct an weighted undirected adjacency graph $G = (V, E)$ with one vertex V_i per data point x_i. We connect vertices V_i and V_j with an edge of weight W_{ij} if x_i is one of the n nearest neighbors of x_j or vice versa. From this, we can determine the so-called graph Laplacian, $\mathbf{L} = \mathbf{W} - \mathbf{D}$, where $\mathbf{D}$ is the diagonal matrix with elements $D_{ii} = \sum_j W_{ji}$.

The graph Laplacian is a positive semidefinite $N \times N$ matrix that satisfies all the properties given above for the continuous Laplacian operator [6]. It follows that if we regard the graph as a mesh on the manifold, the basis determined by the graph Laplacian serves as an approximation to the intrinsic manifold basis. However, the graph analogue is limited to functions that are defined on the graph, not the entire manifold. We denote such functions by the column vector $\mathbf{f} = \langle f_1, ..., f_N \rangle^T$ containing the function image when the domain is limited to the N vertices of the graph. Analogous to the smoothness properties of $\Delta_{\mathcal{M}}$, a measure of smoothness of functions on the graph is given by

$$S_G[\mathbf{f}] = \mathbf{f}^T\mathbf{L}\mathbf{f} = \sum_{i,j} W_{ij}(f_i - f_j)^2. \tag{1}$$

From this functional, the connection of this framework to spectral clustering is highlighted. Consider a logical partition of the graph into two parts $V = V_1 \bigcup V_2, V_1 \bigcap V_2 = \emptyset$. If the graph is faithful to this partition, with vanishing edge weights between V_1 and V_2, then the function that minimizes S_G will reflect the minimum cut (after the constant function, which is ignored). Therefore, the "smoothest" function (i.e. smallest S_G) also acts to separate the two classes.

Therefore, the graph Laplacian approach has dual use. In the case the data provides a sufficiently dense mesh on a manifold, this approach results in an intrinsic basis for the manifold. When the manifold structure is not present, but there still exists a logical partition of the data into two clusters, the graph Laplacian finds functions that perform the clustering. In our experimental section, we will show that success in phoneme classification will require both properties.

2.2. Ordered Intrinsic Basis with Modified Kernelized LPP

Our method of determining the intrinsic basis is based on kernelized locality preserving projections [5], but implements a modified ordering scheme. Consider a training set X composed of l labelled and u unlabelled examples. Using this data, we can form the $(l+u) \times (l+u)$ graph Laplacian matrix, $\mathbf{L}$, as described in the previous section. To determine the intrinsic basis functions, we solve the optimization problem

$$\mathbf{f}^* = \min_{\mathbf{f}^T\mathbf{D}\mathbf{f}=1} \mathbf{f}^T\mathbf{L}\mathbf{f}, \tag{2}$$

the solutions to which satisfy the generalized eigenvalue problem

$$\mathbf{L}\mathbf{f} = \lambda\mathbf{D}\mathbf{f}. \tag{3}$$

Recall that the graph Laplacian method is restricted to functions defined on the graph. This means that the basis this method produces may only be used with the training data that defines the graph. To extend our analysis to functions of novel data points, we implement the theory of reproducing kernel Hilbert spaces (RKHS). Here, we restrict our search for basis functions the RKHS $\mathcal{H}_K$ for some kernel function $K : (X \times X) \to \mathbb{R}$. By the Representer Theorem, we may represent the basis functions, extended out-of-sample, by the form $f(x) = \sum_{i=1}^{l+u} \alpha_i K(x_i, x)$. We can solve for the coefficients of this expansion, $\boldsymbol{\alpha} = \mathbf{K}^+\mathbf{f}$, where $\mathbf{K}^+$ is the Moore-Penrose inverse of the $(l+u) \times (l+u)$ Gram matrix with elements $K_{ij} = K(x_i, x_j)$.

The basis function ordering prescription for kernelized LPP is a simple sort by corresponding eigenvalue, founded on the smoothness priority established in Section 2.1. However, our ultimate goal is binary classification, for which we are concerned primarily with the sign of the projection onto the given component. Therefore, a more relevant measure of classifier component smoothness takes the form

$$m(\mathbf{f}) = \frac{\text{sgn}\,(\mathbf{f})^T\,\mathbf{L}\,\text{sgn}\,(\mathbf{f})}{\text{sgn}\,(\mathbf{f})^T\,\mathbf{D}\,\text{sgn}\,(\mathbf{f})}$$

Given d eigenfunctions $\{\mathbf{f}_1, \ldots, \mathbf{f}_d\}$ of the eigenvalue problem in Equation 3, sorted such that $m_1 = m(\mathbf{f}_1) < m_2 < \cdots < m_d$, we can determine a d-component intrinsic basis defined everywhere, $\{f_1, \ldots, f_d\}$. It follows that the basis functions take the form

$$f_k(x) = \sum_{i=1}^{l+u} \alpha_i^{(k)} K(x_i, x), \tag{4}$$

where $\boldsymbol{\alpha}^{(k)} = \mathbf{K}^+\mathbf{f}_k$.

2.3. Incorporating Labelled Data

Equipped with the first d non-trivial intrinsic basis functions $\{f_1, \ldots, f_d\}$, it is a simple matter to cast the labelled training examples, X_l, into the intrinsic representation. For $x \in X_l$, the projection is determined by $x' = \langle f_1(x), \ldots, f_d(x)\rangle$. We can proceed by estimating a map from the labelled data in this intrinsic representation to their class labels. We limit our study to linear minimum-norm solution maps, but any type of classification scheme may be used here as well, so long as it may function with minimal training data.

Let $\mathbf{F}$ be the $l \times d$ matrix with elements $F_{ij} = f_i(x_j)$, where x_j is the j-th labelled training example. If we denote the column vector of training data labels by $\mathbf{y}_l$, our map is determined by solving the linear system $\mathbf{y}_l = \mathbf{F}\boldsymbol{\beta}$ for $\boldsymbol{\beta} \in \mathbb{R}^d$. In the case of $l \neq d$, this solution must be accomplished using the Moore-Penrose inverse, which is equivalent to the minimum norm solution for $l < d$.

3. Experiments in Phoneme Classification

3.1. The [ɑ]-[æ] Classification Problem in Detail

Our training set consists of 500 randomly chosen training examples of each [ɑ] and [æ] from the TIMIT training dataset. An identically sized and proportioned test set was sampled from the TIMIT test dataset. Each phoneme recording is represented by a 50-dimensional discrete Fourier transform of a ten millisecond window located in the center of the waveform. We compare the performance of two-component LPPSSL ($d = 2$ components; $l = 6$ labelled and $u = 1000 - l$ unlabelled examples) to that of the standard minimum-norm linear classifier ($l = 6$), the equivalent fully-supervised algorithm. For LPPSSL we use a linear kernel and build the adjacency graph with Euclidean distance weights, setting the number of nearest neighbors correspondingly high, at 50. Both algorithms determine a classifying linear hyperplane and neither require model selection.

Since the performance of both algorithms depend heavily on which particular examples are labelled, we repeat the experiment 400 times with different random selections of the labelled set. The computed classifiers for each trial result in the test error histogram shown in Figure 1(a). The fully-supervised minimum-norm linear classifier results in a largely Gaussian distribution, centered about a median test error rate of 24.25%. The semi-supervised performance distribution exhibits a highly skewed Gaussian form with a median test error rate of 11.7% and with 90% of the trials resulting in test error less than the fully-supervised median error of 24.25%. The optimal linear classifier gives a test error rate of 9.7%, as determined with the regularized least squares (RLS) algorithm provided with all 1000 training set labels. Therefore, incorporation of unlabelled data under the manifold/cluster structure assumption reduces test error rates to within 2% of optimal for 50% of the trials when provided only six training examples.

Note that we do not observe an improvement over the optimal test error in any of the trials. This is to be expected since the resulting classifiers are all hyperplanes and any assumption made about the class structure of unlabelled data typically will not be more useful than the class labels themselves. The

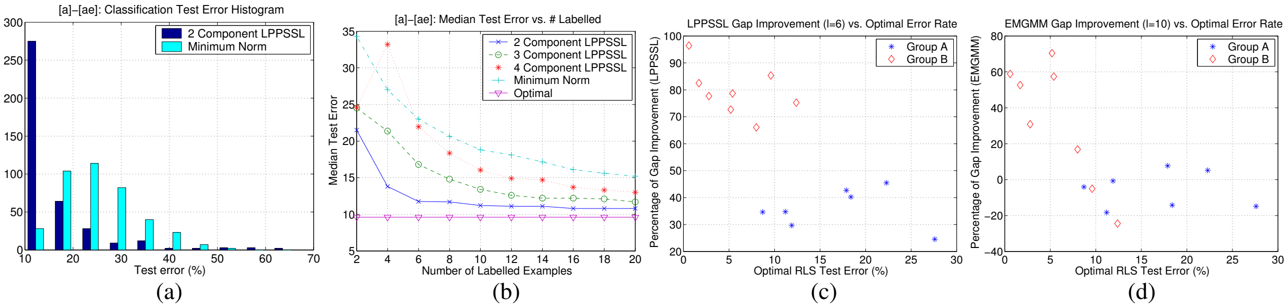

Figure 1: (a) Test error histograms for [ɑ]-[æ] classification ($l = 6$). (b) Test error as a function of number of labelled examples for [ɑ]-[æ] classification. (c) LPPSSL gap improvement, $G(6)$, versus optimal classifier test error for vowel classification problems. (d) EMGMM gap improvement ($l = 10$) versus optimal GMM classifier test error for the fifteen vowel classification problems.

Table 1: Results for fifteen binary classification problems: [optimal test error] / [median min-norm error ($l = 6$)] / [median 2-component semi-supervised LPPSSL error ($l = 6$)] / [$G(6)$]. (All in percent)

	o	ɑ	ɪ	i	æ
ə	27.6/44.3/40.2/24.6	12.4/29.8/16.7/75.2	17.9/30.2/25.0/42.7	5.4/14.3/7.3/78.7	11.2/25.5/20.5/34.7
o	-	18.4/36.9/29.5/40.3	8.0/21.7/12.7/66.1	1.7/7.4/2.7/82.5	8.7/24.3/18.9/34.6
ɑ	-	-	2.8/11.1/4.7/77.7	0.6/3.4/0.7/96.4	9.6/24.3/11.8/85.3
ɪ	-	-	-	22.3/37.7/30.7/45.5	11.9/30.1/24.7/29.7
i	-	-	-	-	5.2/14.9/7.9/72.7

semi-supervised algorithm functions instead to bridge the gap between the optimal solution and the minimum norm solution without requiring more than those six labels. It follows that a useful metric for the semi-supervised performance is the fraction of this gap reduced by the incorporation of unlabelled data:

$$G(l) = \frac{\text{LPPSSL error } (l \text{ labelled}) - \text{Optimal error}}{\text{min-norm error } (l \text{ labelled}) - \text{Optimal error}}. \quad (5)$$

Figure 1(b) shows the [ɑ]-[æ] median test error rate as a function of l for various forms of the fully- and semi-supervised algorithms. Predictably, the two-component semi-supervised test error rate converges toward the optimal as more labels are introduced. However, notice the initially poor performance of the 4-component variant. Here, when providing only two or four labelled examples, the map from these four components to the class labels is overfitted; once the number of labelled examples becomes larger than the number of components, the performance of the three- and four-component variants decreases in the expected fashion. Still, the performance remains slightly worse than the two-component version. This follows from the fact that the third and fourth components do not reflect significant class distinction and these extra components require more data to prevent overfitting.

It follows from Figure 1(b) that the percentage of the min-norm to optimal performance gap (G of Equation 5) bridged by two-component LPPSSL remains roughly constant across all values of l. This constant value serves as a useful measure of overall performance of the algorithm on a given classification problem, and will be used to succinctly characterize performance below.

3.2. Performance Across the Entire Vowel Space

Vowel sounds are traditionally characterized by phoneticians using three production features: horizontal tongue position, vertical tongue position, and lip roundness. Since the six American English vowel phonemes we use in our experiments span this three-dimensional feature space, our performance results are representative of the entire vowel set. Qualitatively speaking, the difficultly of the classification problem scales inversely with distance in production feature space [7]. For example, the [ɑ]-[æ] problem, which differs only in horizontal tongue position, is more difficult than the [ɑ]-[i] problem, which differs in both horizontal and vertical tongue position. Furthermore, lip roundedness is the least distinguishable feature. Therefore, the [ə]-[o] problem is similar in difficulty to the [ɑ]-[o] problem, even though [ə] and [o] are technically further apart in feature space.

Table 1 lists performance results for each of the fifteen binary classification problems involving six vowel phonemes. We find that the baseline separability (i.e., optimal test error) roughly corresponds to the difficultly of the classification problem as indicated by the phonemes' relative separation in production feature space. For instance, the optimal [ɑ]-[i] test error is only 0.6% as compared with 9.6% for [ɑ]-[æ] and 27.6% for [ə]-[o]. Next, consider Figure 1(c), which plots $G(6)$ versus the optimal test error rate for the classification problems. We observe a natural separation of the problems into two groups that emerge from the dual function of the graph Laplacian. Group A pairs have poor separability and thus exhibit minimal clustering of the individual class data. However, the manifold structure still admits a significant gap improvement. Group B pairs are more clearly clustered, resulting in larger gap improvements, each greater than 65%.

3.3. Isolating the Role of Cluster Structure

To further substantiate the significance of LPPSSL manifold properties, we turn to the expectation-maximization algorithm for Gaussian mixture models (EMGMM) as presented in [1]. This algorithm is based solely on the assumption the data is clustered. To impose a fair complexity comparison with minimum-norm methods, we limit the Gaussian mixture model to two components.

We find the EMGMM algorithm applicability in the mini-

Table 2: GMM/EMGMM vowel classification results: [optimal GMM test error] / [median GMM test error ($l = 10$)] / [median semi-supervised EMGMM test error ($l = 10$)] / [Percentage of optimal-GMM performance gap bridged using semi-supervised EMGMM ($l = 10$)]. (All in percent)

	o	ɑ	ɪ	i	æ
ə	40.3/46.4/47.3/-14.9	22.9/37.0/40.5/-24.5	25.9/43.5/42.2/7.7	6.8/36.0/19.3/57.4	20.0/38.3/41.7/-18.3
o	-	37.5/44.9/46.0/-14.2	12.1/40.6/35.8/16.9	4.3/37.7/20.1/52.6	24.0/45.0/45.9/-4.0
ɑ	-	-	5.7/36.7/27.1/30.9	3.3/24.5/12.0/58.9	11.2/45.8/47.5/-5.1
ɪ	-	-	-	31.7/47.3/46.5/5.1	21.9/42.9/43.1/-0.7
i	-	-	-	-	10.7/31.45/16.9/70.4

Table 3: Results for fifteen broad class binary classification problems: [optimal test error] / [median min-norm error ($l = 6$)] / [median 2-comp semi-supervised error ($l = 6$)] / [$G(6)$]. (All in percent)

	Affricates	Fricatives	Nasals	Approximants	Vowels
Stops	10.1/14.8/13.2/33.3	24.7/30.7/28.2/41.7	12.3/21.9/17.9/41.7	21.1/28.4/24.8/49.3	19.5/34.0/28.7/36.9
Affricates	-	17.1/27.2/21.5/56.9	0.9/1.2/1.2/0.0	2.4/3.6/2.9/58.3	0.9/2.9/1.8/55.0
Fricatives	-	-	8.2/11.2/9.4/60.0	10.6/15.0/11.3/85.2	8.4/12.7/9.6/72.0
Nasals	-	-	-	20.6/37.6/28.7/52.6	14.8/34.0/18.2/82.3
Approximants	-	-	-	-	34.9/42.5/39.6/37.7

mal supervision regime to be limited. Table 2 shows the performance of the EMGMM algorithm on the fifteen vowel classification problems outlined above, where we have provided 10 labelled examples. (Note that we were unable to get any consistent improvement by incorporating unlabelled data for $l < 10$.) There exists a clear positive correlation between the fully-supervised optimal GMM and RLS baselines on the fifteen vowel classification test problems, though RLS does consistently perform marginally better. However, for EMGMM we find that the only gains by adding unlabelled data occur in the most separable vowel pairs (e.g. [ɑ]-[i]), namely those with a clear cluster structure. This is consistent with our hypothesis. Also notice that in the case of failure, the incorporation can degrade performance, resulting in negative gap improvement values.

Figure 1(d) shows the EMGMM gap improvement versus the optimal RLS test errors of Figure 1(c), where the same group labels are used. EMGMM fails on the Group A problems, indicating the cluster assumption is not sufficient prior information for successful incorporation of unlabelled data. However, the results indicate that sufficient cluster structure (Group B problems) leads, in most cases, to both moderate EMGMM success and maximal LPPSSL success. Therefore, we conclude that accommodating both manifold and cluster structure is key to success across all vowel classification problems.

3.4. Classification Performance for Broad Classes

Consider six broad classes or phonemes: stops, affricates, fricatives, nasals, approximants (i.e., semivowels and glides), and vowels. We constructed 500-example training and 500-example test sets for each of these six broad classes using the TIMIT dataset. Within each broad class, the individual phonemes are represented according to their occurrence rate in the TIMIT dataset. Table 3 shows the performance of the fully-supervised minimum norm and LPPSSL algorithms on the fifteen broad class classification problems outlined above. Again, we find significant performance improvement from the incorporation of unlabelled data in all but one of the fifteen classification experiments.

When the optimal error rate is low (less than $\sim 15\%$), the fully-supervised minimum-norm solution ($l = 6$) consistently performs better for the broad class problems than for vowels. The superior performance on broad class problems in this regime is a result of a more pronounced cluster structure than that existing in vowel pairs. However, for the LPPSSL results, the performance discrepancy between highly clustered vowel problems and broad class problems is not present. Since LPPSSL relies on both cluster and manifold structure, it is not as sensitive to the separation of clusters, performing equally well on the highly separated broad class clusters and the somewhat closer vowel clusters.

4. Conclusion

We have presented a new manifold-based semi-supervised approach for phoneme classification in the minimal supervision regime. The algorithm's ability to exploit both manifold and cluster structure in vowel and broad class datasets is essential to successful incorporation of unlabelled data. The EMGMM results illustrate that a pure clustering-based semi-supervised approach is not adequate for general success.

5. References

[1] P. Moreno and S. Agarwal, "An experimental study of em-based algorithms for semi-supervised learning in audio classification," in *Proceedings of ICML*, 2003.

[2] A. Jansen and P. Niyogi, "A Geometric Perspective on Speech Sounds," Tech. Rep. TR-2005-08, U. of Chicago, June 2005.

[3] A. Jansen and P. Niyogi, "Intrinsic fourier analysis on the manifold of speech sounds," in *Proceedings of ICASSP*, 2006.

[4] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold Regularization: A Geometric Framework for Learning from Examples," Tech. Rep. TR-2004-06, U. of Chicago, Aug. 2004.

[5] X. He and P. Niyogi, "Locality preserving projections," Tech. Rep. TR-2002-09, Dept. of Computer Science, U. Chicago, 2002.

[6] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[7] Alvin M. Liberman, *Speech: A Special Code*, MIT Press, Cambridge, MA, 1996.