Regularizing the Local Similarity Discriminant Analysis Classifier

Luca Cazzanti
Applied Physics Laboratory
University of Washington
Seattle, USA
luca@apl.washington.edu

Maya R. Gupta
Dept. Electrical Engineering
University of Washington
Seattle, USA
gupta@ee.washington.edu

***Abstract*—We investigate parameter-based and distribution-based approaches to regularizing the generative, similarity-based classifier called local similarity discriminant analysis classifier (local SDA). We argue that regularizing distributions rather than parameters can both increase the model flexibility and decrease estimation variance while retaining the conceptual underpinnings of the local SDA classifier. Experiments with four benchmark similarity-based classification datasets show that the proposed regularization significantly improves classification performance compared to the local SDA classifier, and the distribution-based approach improves performance more consistently than the parameter-based approaches. Also, regularized local SDA can perform significantly better than similarity-based SVM classifiers, particularly on sparse and highly nonmetric similarities.**

***Keywords*-local similarity discriminant analysis; regularized local similarity discriminant analysis;**

I. Similarity-based Classification

Similarity-based classifiers learn from a set of pairwise training similarities, training class labels, and from the similarities between a test sample and the training samples [1]. Similarity-based classifiers are independent of a chosen similarity measure, which is usually problem-dependent and can subsume complex relationships between complex, heterogeneous samples. In this paper, we focus on the problem of designing generative classifiers for similarity-based learning. Here, the goal is to create class-conditional probabilistic models of the given similarities. Generative similarity-based classifiers differ from the standard metric-based generative classifiers, such as quadratic discriminant analysis and Gaussian mixture models, because the modeled quantity is the pairwise similarities between the samples rather than the numerical feature vectors that describe the samples. Producing class probabilities is important in many practical systems where there may be skewed class priors or asymmetric misclassification costs, or where probabilities are required as an input to the next component in the system or to fuse with probabilistic information about the class label derived from other sources.

Recently, an effective generative classifier for similarity-based learning called similarity discriminant analysis (SDA) and a local version (local SDA) were proposed [2], [3]. We review local SDA in Section 3, and discuss how this classifier can fail. In Section 4, we follow our analysis with a discussion of several regularization strategies for local SDA and with the main contribution of this paper: that appropriate regularization can both make the SDA model more flexible and lower the estimation variance. Experiments in Section 5 show that the proposed regularized local SDA improves on local SDA and can outperform other state-of-the-art similarity-based classifiers.

Previous research on generative classifiers for similarity-based learning treated the n-vector of similarities between any sample and the n training samples as a feature vector, and in this way created a Euclidean space based on the training similarities. Standard generative classifiers were then trained in this space, such as regularized linear or quadratic discriminant analysis [4], [5]. These classifiers have the drawback that their generative models grows as $O(n)$ and $O(n^2)$, and are arguably difficult to interpret.

Another vein of related research considers generative models for random graphs or networks [6]. This research models the distribution of graphs exponentially, and could be directly applied for cases of similarity-based learning where similarities only took on binary values. For the general case of similarity-based learning, how to adapt this type of generative model is an open question.

II. Background on Similarity-based Learning

In this section we discuss some applications in which similarity-based learning arises, and review other approaches to similarity-based classification; see [1] for a more thorough review of related work.

A. Similarity-based Problems

Similarities that stem from heterogeneous data often do not manifest the properties of inner products. These non-metric similarities arise naturally in fields such as genomics, natural language processing, human commerce, computer vision, and psychology.

Laub et al. [7] demonstrate that indefinite similarities can encode useful information in their non-metricity, and contend "in the realm of human similarity judgments, one may not

Work supported by the U.S. Office of Naval Research

speak of artifact or erroneous judgements with respect to an Euclidean norm. On the contrary: having an Euclidean norm is rather the exception.”

In this paper we will show results on experiments for four different similarities, which we detail here to illustrate the diversity of classification problems for which similarity-based classifiers are a natural fit. These datasets stem from human experience and are available from the Similarity-based Learning Repository [1].

The Amazon problem is to classify books as fiction or non-fiction, where the similarity between two books is the symmetrization of the percentage of customers who bought one books after viewing the other book. There are 96 samples in this dataset, 36 from class Non-fiction, and 60 from class Fiction. This similarity function strongly violates the triangle inequality. This dataset is also especially interesting because this similarity strongly violates the minimality property that says a sample should be maximally similar to itself, because customers often buy a different book if they first view a poorly-reviewed book. Fig. 1 shows the 96×96 similarity matrix and its spectra.

The Aural Sonar problem is to distinguish 50 target sonar signals from 50 clutter sonar signals. Listeners perceptually evaluated the similarity between two sonar signals on a scale from 1 to 5. The pairwise similarities are the sum of the evaluations from two listeners, resulting in a perceptual similarity from 2 to 10 [8]. This dataset is interesting because perceptual similarities are often non-metric.

The Patrol problem is to classify 241 people into one of 8 patrol units based on who people claimed was in their unit when asked to name five people in their unit [9]. Like the Amazon dataset, this is a sparse dataset, as most of the similarities are zero.

The Voting problem is to classify 435 representatives into two political parties based on their votes [10]. The categorical feature vector of yes/no/abstain votes was converted into pairwise similarities using the value difference metric, which is a (dis)similarity designed to be useful for classification [11]. The voting similarity is almost metric.

B. Other Similarity-based Classifiers

Besides generative classifiers, there are three major approaches to similarity-based classification: nearest neighbor methods, treating similarities as features, and treating similarities as kernels. Treating similarities as features means using the similarities to the n training samples (or some subset thereof) as a feature vector, and then classifying with any standard metric learning algorithm. In our experiments, we represent this approach with a support vector machine (SVM) applied (with a linear kernel) to the similarities-as-features [12]. Treating similarities as kernels entails approximating the similarity training matrix by a symmetric positive definite matrix which can be used as a kernel in an SVM. Three popular ways to approximate the similarity matrix are to set any negative eigenvalues to zero (clip), to flip any negative eigenvalues so they become positive (flip), and to add the identity matrix scaled by the minimum (negative) eigenvalue to the original similarity matrix (shift) [13]. Clipping the spectrum produces the nearest positive semidefinite matrix in terms of the Frobenius norm. Flipping the eigenvalues is

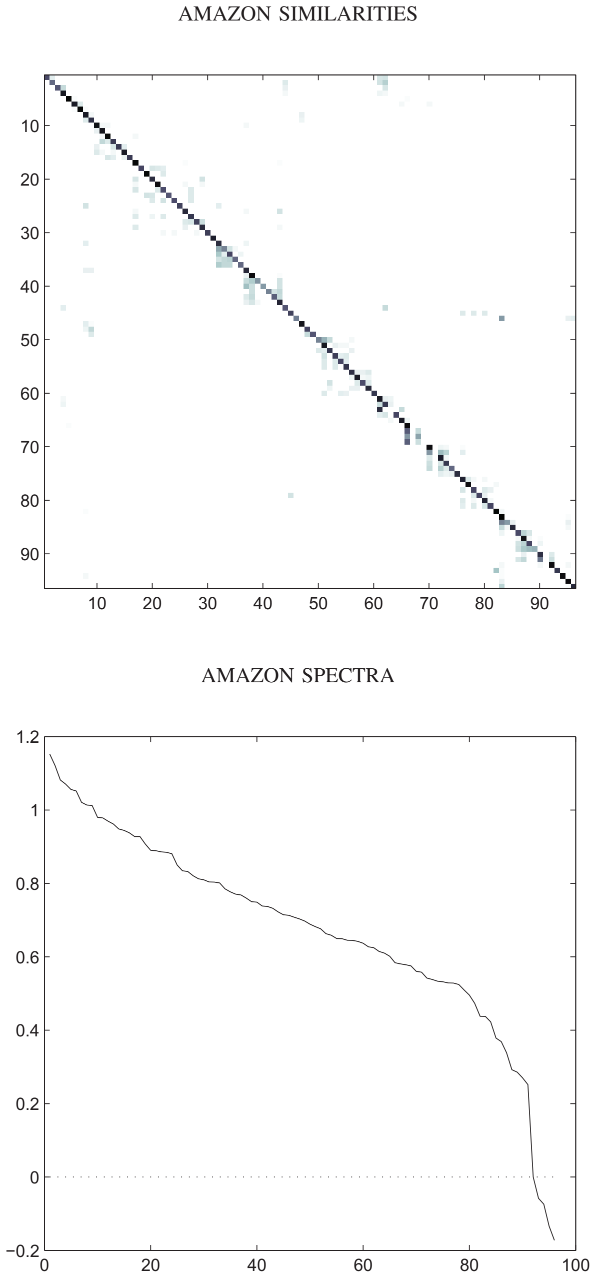

Fig. 1. **Top:** The 96×96 similarity matrix is shown for the samples in the Amazon dataset. Darker color corresponds to greater similarity. One sees from the diagonal that the Amazon samples are not necessarily most similar to themselves. Also, the Amazon similarities are mostly zero (shown as white). **Bottom:** The eigenvalues for the Amazon similarity matrix are shown; Amazon has the largest relative-magnitude negative eigenvalues out of the five datasets considered in the experiments.

[1] idl.ee.washington.edu/similaritylearning

similar to treating SS^T as a kernel, and thus behaves similarly to using the similarities-as-features. Shifting the spectrum has the property that only the self-similarities are changed [14].

III. Review and Analysis of Local SDA

In this section we briefly review the local SDA classifier and motivate the need for regularization.

A. Review of Local SDA

Assume that the test and training samples belong to an abstract space of possible samples $\mathcal{B}$, such as the set of all books. Let $X \in \mathcal{B}$ be a random test sample with random class label $Y \in \{1, 2, \ldots, G\}$, and let $x \in \mathcal{B}$ denote the realization of X. Also assume that one can evaluate a relevant similarity function $s : \mathcal{B} \times \mathcal{B} \to \Omega$, where $\Omega \subset \mathcal{R}$ is assumed to be a finite discrete space without loss of generality. Let $\mathcal{X} \subset \mathcal{B}$ be the set of n training samples.

The local SDA classifier follows from the standard Bayes classifier by making the fundamental assumption that all the information about X's class label depends only on local similarity statistics. Previous work considered different variants [15], but here we restrict attention to the local centroid variant that seems to be most effective [2].

Given a test sample x, the local centroid $\mu_h(x)$ for class h is defined as the local training sample with maximum sum-similarity to its class:

$$\mu_h(x) \text{ solves } \arg \max_{a \in \mathcal{X}_h \cap \mathcal{N}(x)} \sum_{z \in \mathcal{X}_h \cap \mathcal{N}(x)} s(z, a), \quad (1)$$

where $\mathcal{X}_h \subset \mathcal{X}$ is the subset of training samples from class h and $\mathcal{N}(x)$ is the neighborhood of x defined as its k nearest (most similar) training samples. Then, the classification rule for the local SDA classifier is to classify a test sample x as the class $\hat{y}$ that minimizes the expected misclassification costs[2]:

$$\arg \min_{f=1,\ldots,G} \sum_{g=1}^{G} C(f, g) \left(\prod_{h=1}^{G} \gamma_{gh} e^{\lambda_{gh} s(x,\mu_h)} \right) P(g), \quad (2)$$

where $C(f, g)$ is the cost of classifying as class f when the true class is g, $P(g)$ is the a priori probability of class g, and $P(x|Y = g)$ is modeled as $\prod_h P(s(x, \mu_h)|Y = g) = \prod_h \gamma_{gh} e^{\lambda_{gh} s(x,\mu_h)}$, that is, the product of class-conditional probabilities of the similarity of x to the centroid μ_h of class h, given that x is in class g.

The parameters $\{\lambda_{gh}\}$ are determined by numerical minimization under the method-of-moments constraint that the expected value of the similarity be equal to the observed average similarity,

$$E_{P(s(x,\mu_h)|Y=g)}[s(X, \mu_h)] = \frac{\sum_{z \in \mathcal{X}_g \cap \mathcal{N}(x)} s(z, \mu_h)}{|\mathcal{X}_g \cap \mathcal{N}(x)|}, \quad (3)$$

and the parameters $\{\gamma_{gh}\}$ are determined by normalization.

In summary, the SDA model can be interpreted as the maximum likelihood exponential model with independence assumptions, or equivalently as the maximum entropy distribution that satisfies the set of mean constraints given by (3).

B. Analysis of Local SDA

We have found experimentally that the local class-conditional estimates have high variance, and that often the local SDA classifier must degenerate to a simpler model. Consider the case that a neighborhood $\mathcal{N}(x)$ contains only one training sample from any class g, then the class-conditional pmfs for that class would be a Kronecker delta with all its probability concentrated on the self-similarity of the single neighborhood training sample. Unless the test sample happened to have that exact similarity, the likelihood of the test sample's similarities would be considered zero for that class. Similar difficulties can arise if there are only a few neighbors from a class, or in general if the resulting local model is degenerate. In fact, our analysis suggests that this degeneracy problem occurs often, and that how it is handled can significantly change the performance of local SDA.

In the original local SDA paper [2], the degeneracy problem was handled by reverting to a local nearest centroid classifier [16] for all the classes if any class had exactly one neighbor. This is suboptimal for multiple reasons: (1) degeneracies can happen even if there is more than one neighbor; (2) for classification problems with many classes it is likely that some class will only have one neighbor in a test sample's neighborhood; (3) reverting to a local nearest centroid classifier does not produce probability estimates; and (4) reverting to a local nearest centroid classifier creates a conceptual and empirical discontinuity as the neighborhood size changes.

Another approach to cope with the local scarcity of samples from one or more classes is to increase the size of the neighborhood k until satisfactory estimates of each class-conditional distribution can be achieved. This approach decreases the localized quality of local SDA, potentially increasing the model bias because a larger neighborhood size must be used.

IV. Regularized Local SDA

We propose regularizing the local SDA classifier as a way to solve the degenerate model problem described in Sec III-B, while retaining both the localized and generative characteristics of local SDA. We discuss three methods for regularizing the local SDA classifier: regularizing the parameters λ_{gh}, regularizing the mean similarity constraint (3), and regularizing the class-conditional distributions $P(s(x, \mu_h)|Y = g)$. We argue that this third option is most consistent with the model intuition and affords the best bias-variance trade-off.

A. Regularizing the Model Parameters

One of the first regularized generative models was Friedman's proposal to regularize the covariance matrix in quadratic discriminant analysis (QDA) [17]. A straightforward generalization of Friedman's success with regularized QDA would

[2] We drop the cumbersome notation $\mu_h(x)$ from the rest of the paper in favor of μ_h, but it should be understood that the class centroids are determined from the neighborhood $\mathcal{N}(x)$, and thus depend on the test sample x.

suggest regularizing the model parameter λ_{gh}, by linearly combining it with a more stable analogous parameter $\tilde{\lambda}_{gh}$:

$$\hat{\lambda}_{gh} = (1-\alpha)\lambda_{gh} + \alpha\tilde{\lambda}_{gh}, \tag{4}$$

where the parameter $0 \leq \alpha \leq 1$ controls the weight of the regularization and can be chosen by cross-validation. The parameter $\tilde{\lambda}_{gh}$ could for example be the model parameter for the corresponding global SDA model (that is, for $k = n$), or could be the model parameter averaged over classes:

$$\tilde{\lambda}_{gh} = \tilde{\lambda} = \frac{1}{G^2}\sum_{g=1}^{G}\sum_{h=1}^{G}\lambda_{gh}.$$

This regularization will generally result in flatter exponential models. The regularized parameter $\hat{\lambda}_{gh}$ is the solution to some mean constraint

$$E_{\hat{P}(s(x,\mu_h)|Y=g)}[s(X,\mu_h)] = c, \tag{5}$$

but because the similarity domain is discrete and finite, there is no straightforward relationship between the c in (5) and the unregularized empirical average in the constraint (3). Thus, the intuitive connection between $\hat{\lambda}_{gh}$ and the mean similarity is lost, and we find it difficult to interpret this regularization or relate it back to the maximum entropy framework used to justify local SDA.

B. Regularizing the Empirical Mean

A more interpretable regularization results from regularizing the empirical class-conditional mean in (3) and solving for the exponential model parameter $\hat{\lambda}_{gh}$ that satisfies:

$$\begin{aligned} & E_{\hat{P}(s(x,\mu_h)|Y=g)}[s(x,\mu_h)] \\ = \quad & (1-\alpha)\frac{\sum_{z\in\mathcal{X}_g\cap\mathcal{N}(x)} s(z,\mu_h)}{|\mathcal{X}_g\cap\mathcal{N}(x)|} + \alpha\frac{\sum_{z\in\mathcal{X}_g} s(z,\tilde{\mu}_h)}{|\mathcal{X}_g|}, \end{aligned}$$

where $\tilde{\mu}_h$ is the global centroid for the hth class (defined by (1) for $\mathcal{N}(x) = \mathcal{X}$).

This strategy shifts the local mean similarity toward the global mean similarity, and the resulting pmf is shifted correspondingly, generally towards a flatter exponential. The motivation for this type of regularization is that linearly combining the local and global means before estimating the parameter moderates the classification variance caused by the randomness of choosing only a local neighborhood of training samples.

We note a drawback to this regularization approach. A good regularizer should be a more stable version of what is ill-posed. But, we expect the global average similarity to the global centroid $s(z,\tilde{\mu}_h)$ to differ qualitatively from the local average similarity that it is regularizing in (6).

C. Regularizing the Class-conditional Probabilities

The approaches in the previous two sections regularize the parameters that control the class-conditional pmfs $P(s(x,\mu_h)|Y=g)$ by either directly regularizing the exponents $\{\lambda_{gh}\}$, or indirectly regularizing them by regularizing the mean constraints. In contrast, here we propose to regularize the local class-conditional pmfs themselves by linearly combining them with the average of the local class-conditional pmfs computed from the training set. This regularization should reduce estimation variance, but also enlarges the model space to include more than exponential distributions, thus increasing the model flexibility.

Let $z \in \mathcal{X}$ be a training sample, and let $\mathcal{N}(z) \subset \mathcal{X}$ be its neighborhood consisting of its k most similar samples from $\mathcal{X}$. Let training sample's z's class h local centroid $\mu_h(z)$ be computed from $\mathcal{N}(z)$, that is

$$\mu_h(z) \text{ solves } \arg\max_{a\in\mathcal{X}_h\cap\mathcal{N}(z)} \sum_{q\in\mathcal{X}_h\cap\mathcal{N}(z)} s(q,a).$$

Denote by $P_z(s(q,\mu_h(z))|Y=g)$ the gth-hth local class-conditional pmf computed from z's neighborhood such that is $P_z(s(q,\mu_h(z))|Y=g)$ is the class-conditional pmf determined by solving the mean constraint,

$$\begin{aligned} & E_{P(s(q,\mu_h(z))|Y=g)}[s(Q,\mu_h(z))] \\ & \quad = \frac{\sum_{q\in\mathcal{X}_g\cap\mathcal{N}(z)} s(q,\mu_h(z))}{|\mathcal{X}_g\cap\mathcal{N}(z)|}. \end{aligned} \tag{6}$$

Then to each training sample z there corresponds a local class-conditional pmf $P_z(q,\mu_h(z))|Y=g)$ for $g = 1,\ldots,G$ and $h = 1,\ldots,G$, and repeating the process for each of the n training samples creates a set of n pmfs $\{P_z(q,\mu_h(z))|Y = g)\}$ for each of the possible G^2 choices of g and h. We average these training local class-conditional pmfs for each choice of g and h:

$$P_{ave}(s_h|g) \overset{\triangle}{=} \frac{1}{|\mathcal{X}|}\sum_{z\in\mathcal{X}} P_z(s(q,\mu_h(z))|Y=g). \tag{7}$$

Each average local class-conditional pmf P_{ave} is linearly combined with the corresponding test sample's estimated local class-conditional pmf to produce the regularized local class-conditional pmfs used in the classifier. That is, given a test sample x and its neighborhood $\mathcal{N}(x)$ of size k, the G^2 local pmfs are each regularized as

$$\begin{aligned} & \hat{P}(s(x,\mu_h)|Y=g) \\ = \quad & (1-\alpha)P(s(x,\mu_h)|Y=g) + \alpha P_{ave}(s_h|g) \end{aligned} \tag{8}$$

The proposed *marginal-regularized local SDA* classification rule then takes the product of the assumed-independent regularized marginals:

$$\arg\min_{f=1,\ldots,G}\sum_{g=1}^{G} C(f,g)\left(\prod_{h=1}^{G}\hat{P}(s(x,\mu_h)|Y=g)\right)P(g). \tag{9}$$

We recently proposed an alternative distribution-based method that regularizes based on the local joint pmfs [18]. That is, we first regularize each marginal class-conditional pmf, and multiply the regularized marginals to form the G class-conditional joint pmfs used in (9). For notational simplicity, denote by $\mathcal{T}(x) \overset{\triangle}{=} \{s(x,\mu_1),\ldots,s(x,\mu_G)\}$, the set of similarities of test sample x to the class centroids. Also denote each class-conditional joint pmf and its regularized counterpart by

$$\begin{aligned} P(\mathcal{T}(x)|Y=g) &\triangleq \prod_{h=1}^{G} \gamma_{gh} \exp(\lambda_{gh} s(x,\mu_h)) \\ \tilde{P}(\mathcal{T}(x)|Y=g) &\triangleq \prod_{h=1}^{G} P_{ave}(s_h|g). \end{aligned}$$

Then we compute regularized class-conditional joint pmfs as

$$\begin{aligned} \hat{P}(\mathcal{T}(x),g) = \quad & (1-\alpha)P(\mathcal{T}(x)|Y=g)P(g) + \\ & \alpha\tilde{P}(\mathcal{T}(x)|Y=g)\tilde{P}(g), \end{aligned} \tag{10}$$

where $P(g)$ and $\tilde{P}(g)$ are the g class prior probabilities estimated from $\mathcal{N}(x)$ and from $\mathcal{X}$ respectively. Then the proposed *joint-regularized local SDA* classification rule is

$$\arg\min_{f=1,\ldots,G} \sum_{g=1}^{G} C(f,g)\hat{P}(\mathcal{T}(x),g). \tag{11}$$

The proposed approach of regularizing the pmfs themselves has several desirable properties. First, any degenerate Kronecker delta functions that arise as solutions for a local pmf $P(s(x,\mu_h)|Y=g)$ are regularized by a smoother pmf $P_{ave}(s_h|g)$. The regularizing pmf $P_{ave}(s_h|g)$ is itself smooth because it is an average of many local exponential pmfs computed from the training set, and any Kronecker delta functions that arise as solutions for any particular $P_z(s(q,\mu_h(z))|Y=g)$ are averaged with the rest of the local training pmfs.

A second desirable property of regularizing the local class-conditional pmfs is the increased modeling flexibility. In general, each $P_{ave}(s_h|g)$ is not exponential, but rather a weighted sums of class-conditional exponential functions over a discrete similarity domain. In fact, because the SDA framework allows both positive and negative values for $\{\lambda_{gh}\}$, the regularizing pmfs can flexibly model complicated distributions of the class-conditional similarities.

A third desirable property of the proposed regularization is that it regularizes using analogous quantities: the test sample's local class-conditional pmf is regularized by other samples' local class-conditional pmfs.

Lastly, we argue that the action of regularizing the local test pmfs with the corresponding local training sample pmfs is easy to interpret in terms of the effect of choices of the regularization parameter α, which makes it easier to decide cross-validation choices and to interpret the cross-validated regularization parameter.

V. Experiments

We compare the marginal-regularized and joint-regularized local SDA to the original local SDA, and to five other state-of-the-art similarity-based classifiers.

Four datasets were partitioned 20 times into disjoint benchmark partitions of 80% training samples and 20% test samples. For each of the 20 partitions of each dataset we chose parameters using ten-fold cross-validation for each of the classifiers shown in Table I. Cross-validation parameter sets were based on recommendations in previously published papers and popular usage. For k-NN and the local SDA classifiers, the choice of neighborhood sizes was $k \in \{1,2,3,\ldots,16,32,64,\min(n,128)\}$. For regularized local SDA, the choices for the convex regularizing parameter were $\alpha \in \{10^{-6},10^{-3},0.01,0.1,0.5,0.9\}$. For all four SVMs, the standard C parameter choices were $10^{-3},10^{-2},\ldots,10^{5}$. Multi-class implementations of the SVM classifiers used $\binom{n}{2}$ pairwise classifiers.

A. Classification Results

Results are shown in Table I, averaged over the 20 randomized partitions. For each dataset, the best classifier and the classifiers not statistically significantly worse are in bold, where the significance was evaluated with the one-sided Wilcoxon signed rank test ($p = 0.05$).

For all four datasets, at least one of the the regularized local SDA classifiers provides better performance than the local SDA classifier, and for three datasets the performance improvement is statistically significant. Regularizing either the joint or the marginal local SDA pmfs improves classification across different data sets more consistently than regularizing the local SDA parameters. The most significant gains are on the Amazon and Aural Sonar datasets: regularizing the joint distributions produces a 19% gain on Amazon and a 8% gain on Aural Sonar over the original local SDA, and regularizing the marginal distributions produces a 9.5% gain on Amazon and 32% gain on Aural Sonar. Compared to other state-of-the-art similarity-based classifiers, regularized SDA performs better for three datasets, and significantly better than the SVM classifiers on two datasets. The performance gap is biggest for sparse similarities like those encountered with the Amazon and Patrol datasets, for which regularized local SDA seems to provide a competitive advantage.

VI. Discussion and Hypotheses

We investigated five regularization approaches for the generative, similarity-based local SDA classifier. We discussed why regularizing the exponential parameters or the mean constraints may be the strictest generalization of other successful regularization methods like regularized QDA, but provide a suboptimal bias-variance trade-off for the local SDA classifier. We argued that regularizing the test sample's exponential local class-conditional pmfs with an average of local training class-conditional pmfs can provide a better bias-variance trade-off than regularizing the parameters. Regularizing local pmfs with average local pmfs also provides a more straightforward interpretation of the regularization process, because conceptually analogous quantities are combined.

Experimental results demonstrated that this proposed regularization consistently and sometimes significantly improves the classification performance of the local SDA classifier, and can perform better than similarity-based SVM classifiers. In particular, we hypothesize that the regularized local SDA classifier will perform best relative to global kernel-based

TABLE I

AVERAGE (STANDARD DEVIATION) OF THE PERCENT TEST ERROR OVER 20 RANDOM TEST/TRAIN SPLITS. FOR EACH DATASET, THE BEST CLASSIFIER AND THE CLASSIFIERS NOT STATISTICALLY SIGNIFICANTLY WORSE ARE IN BOLD.

	Amazon (2 classes)	Aural Sonar (2 classes)	Patrol (8 classes)	Voting (2 classes)
Local SDA	11.05 *(7.61)*	17.75 *(7.66)*	**11.77** *(4.62)*	6.38 *(2.07)*
Local SDA *(joint-regularized)*	**8.95** *(5.79)*	16.25 *(5.67)*	**11.35** *(4.63)*	**5.40** *(1.63)*
Local SDA *(marginal-regularized)*	**10.00** *(6.21)*	**12.00** *(6.20)*	11.98 *(4.36)*	**5.46** *(2.05)*
Local SDA *(global $\tilde{\lambda}_{gh}$-regularized)*	**10.00** *(7.42)*	17.25 *(5.36)*	11.87 *(4.42)*	5.63 *(1.89)*
Local SDA *(average $\bar{\lambda}$-regularized)*	**10.26** *(7.15)*	18.25 *(5.54)*	11.87 *(4.42)*	**5.46** *(2.12)*
Local SDA *(mean constraint-regularized)*	11.32 *(8.69)*	17.00 *(4.85)*	11.98 *(4.70)*	5.06 *(1.83)*
k-NN	**9.47** *(6.57)*	17.00 *(7.65)*	11.88 *(4.42)*	5.80 *(1.83)*
SVM w/ clipped spectrum	12.37 *(7.68)*	**13.00** *(5.34)*	38.75 *(4.81)*	**4.89** *(2.05)*
SVM w/ flipped spectrum	20.79 *(10.97)*	**13.25** *(5.31)*	47.29 *(5.90)*	**4.94** *(2.03)*
SVM w/ shifted spectrum	15.53 *(13.05)*	**14.00** *(5.61)*	40.83 *(5.37)*	**5.17** *(1.87)*
SVM on similarities as features	16.05 *(11.59)*	**14.25** *(6.94)*	42.19 *(5.85)*	5.40 *(2.03)*

methods when the similarities are sparse and strongly nonmetric.

For local SDA and its regularized variants, classifying a test sample x requires estimating a new local class centroid $\mu_h(x)$, which is determined as the neighbor from class h whose sum-similarity to other neighbors from the same class is maximum. In practice the cross-validated neighborhood sizes are relatively small so that the concomitant cost of summing a few similarity values for each test sample is negligible.

Standard generative classifiers such as QDA estimate G class-conditional models. In contrast, SDA-type classifiers estimate G^2 class-conditional models by solving G^2 mean-constrained function minimization problems. Regularized local SDA further requires solving one set of G^2 constraints for each training sample, and the required classifier training time is proportional to nG^2. For large datasets the training time can become impractical, although we emphasize that training is done only once per dataset and the estimated regularizing pmfs are stored in memory, so that classifying each test sample is an almost instantaneous operation.

An approach to lowering the training costs of our proposed distribution-based SDA regularization schemes might be to reqularize the local SDA pmfs with an exponential pmf fitted to the average histogram of local similarity counts. This approach would require solving only one set of G^2 mean-constrained function minimizations for each dataset during the training phase and would rely on counting the occurrences of the local similarity values, which is much faster than solving many optimization problems.

Here, we used as a regularizer the simple average of the training sample pmfs. An extension to this work might adaptively regularize, for example by weighting each training sample's pmf in the regularizer based on the relative similarity of each training sample to that particular test sample. Yet another approach might be to adopt Bayesian methods to estimate the local SDA pmfs, for example by enlarging the space of possible pmfs to include the multinomial and Dirichlet families. Characterizing the trade-offs between the classification performance, the model flexibility, and the computational efficiency of these possible alternatives remains an open area of research.

REFERENCES

[1] Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti, "Similarity-based classification: Concepts and algorithms," *Journal of Machine Learning Research*, vol. 10, pp. 747–776, March 2009.

[2] L. Cazzanti and M. R. Gupta, "Local similarity discriminant analysis," in *Proc. Intl. Conf. Machine Learning*, 2007.

[3] L. Cazzanti, M. R. Gupta, and A. J. Koppal, "Generative models for similarity-based classification," *Pattern Recognition*, vol. 41, no. 7, pp. 2289–2297, July 2008.

[4] E. Pekalska, P. Paclíc, and R. P. W. Duin, "A generalized kernel approach to dissimilarity-based classification," *Journal of Machine Learning Research*, pp. 175–211, 2001.

[5] E. Pekalska and R. P. W. Duin, "Dissimilarity representations allow for building good classifiers," *Pattern Recognition Letters*, vol. 23, no. 8, pp. 943–956, June 2002.

[6] M. Handcock, D. R. Hunter, and S. Goodreau, "Goodness of fit of social network models," in *Journal American Statistical Association*, vol. 103, no. 1, 2008, pp. 248–258.

[7] J. Laub, V. Roth, J. M. Buhmann, and K. Müller, "On the information and representation of non-Euclidean pairwise data," *Pattern Recognition*, vol. 39, pp. 1815–1826, 2006.

[8] S. Philips, J. Pitton, and L. Atlas, "Perceptual feature identification for active sonar echoes," in *Proc. of the 2006 IEEE OCEANS Conf.*, 2006.

[9] J. E. Driskell and T. McDonald, "Identification of incomplete networks," *Florida Maxima Corporation Technical Report*, no. 08–01, 2008.

[10] A. Asuncion and D. J. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html

[11] C. Stanfill and D. Waltz, "Toward memory-based reasoning," *Communications of the ACM*, vol. 29, no. 12, pp. 1213–1228, December 1986.

[12] T. Graepel, R. Herbrich, P. Bollmann-Sdorra, and K. Obermayer, "Classification on pairwise proximity data," in *Advances in Neural Information Processing Systems*, vol. 11, 1998, pp. 438–444.

[13] G. Wu, E. Y. Chang, and Z. Zhang, "An analysis of transformation on non-positive semidefinite similarity matrix for kernel machines," University of California, Santa Barbara, Tech. Rep., March 2005.

[14] V. Roth, J. Laub, M. Kawanabe, and J. M. Buhmann, "Optimal cluster preserving embedding of nonmetric proximity data," *IEEE Trans. Pattern Anal. and Machine Intel.*, vol. 25, no. 12, pp. 1540–1551, Dec. 2003.

[15] L. Cazzanti, "Generative models for similarity-based classification," Ph.D. dissertation, Department of Electrical Engineerng, University of Washington, 2007.

[16] Y. Mitani and Y. Hamamoto, "Classifier design based on the use of nearest neighbor samples," *Proc. of the Intl. Conf. on Pattern Recognition*, pp. 769–772, 2000.

[17] J. H. Friedman, "Regularized discriminant analysis," *Journal American Statistical Association*, vol. 84, no. 405, pp. 165–175, 1989.

[18] L. Cazzanti and M. R. Gupta, "Fusing similarities and Euclidean features with generative classifiers," in *Proc. 12th Intl. Conf. Information Fusion*, July 2009.