Bayesian and Pairwise Local Similarity Discriminant Analysis

Peter Sadowski
Dept. Electrical Engineering
University of Washington
Seattle, USA
psadowsk@ee.washington.edu

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 three extensions to the generative similarity-based classifier called local similarity discriminant analysis (local SDA): a Bayesian approach to estimating the pmfs based on the assumption that similarities are multinomially distributed and on the Dirichlet prior distribution; a pairwise-similarity formulation of local SDA that accounts for all local pairwise similarities to estimate the pmfs; a combined Bayesian pairwise-similarity approach. We discuss how the proposed extensions afford more modeling flexibility than standard local SDA and less cumbersome model training than previously-published local SDA regularization strategies. Experiments with five benchmark similarity-based classification datasets show that the increased modeling flexibility and lighter computational burden of the proposed extensions are coupled with the good classification performance of the local SDA classification paradigm.**

***Index Terms*—similarity-based classification; discriminant analysis; Bayesian; prototype; Dirichlet distribution;**

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 similarity 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) was proposed [2], followed by a local version (local SDA) [3], and a regularized local version [4]. We review the standard local SDA classifier in Section II, and discuss its limitations. In Section III, we introduce the first contribution of this paper: a Bayesian framework for estimating the local SDA class-conditional pmfs based on a the assumption that the similarities are multinomially distributed and on a Dirichlet-distributed prior. In Section IV we introduce the second contribution of this paper: a pairwise local SDA classifier which endows local SDA with increased flexibility and robustness. We also discuss a combined pairwise, Bayesian approach to modeling the similarities. Experiments in Section V show that the proposed, more flexible approaches do not impair the local SDA performance, often produce better results, and are competitive with other state-of-the-art similarity-based classifiers.

There has been other research into generative classifiers for similarity-based learning that treats the n-vector of similarities between any sample and the n training samples as a feature vector, and then applies standard generative classifiers to that feature space. These classifiers, such as regularized linear or quadratic discriminant analysis [5], [6] have the drawback that their generative models grows as $O(n)$ and $O(n^2)$, and are arguably difficult to interpret.

Other related research considers generative models for random graphs or networks [7]. They model the a graph as being drawn from an exponential distribution, and theoretically we believe this model could be applied for cases of similarity-based learning where similarities only take on binary values, but no such experiments have been done. For the general case of similarity-based learning, how to adapt this type of generative model is an open question.

Besides generative classifiers, there are three major approaches to similarity-based classification: nearest neighbor methods, treating similarities as features, and treating similarities as kernels; for a more thorough review of these methods see [1]. 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, as done in the related generative classifier work noted above. In our experiments, we represent this approach with a local support vector machine (SVM) applied with a linear kernel to the similarities-as-features [8]. Treating similarities as kernels entails approximating the similarity training matrix S by a symmetric positive definite matrix which can be used as a kernel in a 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) [9]. Clipping the spectrum produces the nearest positive semidefinite matrix in terms of the Frobenius norm. As argued in [1], flipping the eigenvalues has a similar effect as using 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 [10]. Finally, local discriminative classifiers (SVM-KNN) produce local SVMs from the k-nearest (most similar) neighbors of a test sample [11]. Any of the eignevalue manipulation techinques can be applied to the local pairwise similarity matrices to produce local kernels.

II. Review of Local SDA

In this section we review the standard local SDA classifier, discuss its limitations and previous regularization strategies, and motivate the contributions of this paper.

Assume that the test and training samples belong to an abstract space of possible samples $\mathcal{B}$. 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 $\Omega \subset \mathbb{R}$ is assumed to be a finite discrete space without loss of generality, and $r = |\Omega|$. Let $\mathcal{X} \subset \mathcal{B}$ be the set of n training samples, and $\mathcal{N}(x) \subset \mathcal{X}$ be the neighborhood of a test sample x, defined as its k-nearest (most similar) training samples. Also, let $\mathcal{N}_g(x) \subset \mathcal{N}(x)$ be the subset of x's neighbors that belong to class g.

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 a set of local similarity statistics, $\mathcal{T}(x) = \{T_1(x), T_2(x), \ldots T_M(x)\}$. The classification rule for local SDA is is to classify a test sample x as the class $\hat{y}$ that minimizes the expected misclassification costs

$$\arg \min_{f=1,\ldots,G} \sum_{g=1}^{G} C(f,g) P(\mathcal{T}(x)|Y=g) P(g), \qquad (1)$$

where $C(f,g)$ is the cost of classifying as class f when the true class is g, $P(x|Y=g)$ is modeled as $P(\mathcal{T}(x)|Y=g)$, and $P(g)$ is the a priori probability of class g.

Many choices are possible for $\mathcal{T}(x)$ [12], but of the choices we had previously considered, the best performing was $s(x, \mu_h(x))$, the similarity between the sample x and its local class centroid $\mu_h(x)$ [3], where $\mu_h(x) \in \mathcal{N}_h(x)$ is the local training sample from class h with maximum sum-similarity to all other training samples from its same class,

$$\mu_h(x) = \arg \max_{v \in \mathcal{N}_h(x)} \sum_{a \in \mathcal{N}_h(x)} s(a, v).$$

.

Given these similarities to centroids as the set of similarity statistics $\mathcal{T}(x) = \{s(x, \mu_h)\}^1$,where $h = 1 \ldots G$, the standard local SDA classifier models the probability of the similarity of x to the centroid μ_h as a discrete exponential marginal pmfs, conditioned on x being from class g. The joint class-conditional pmfs is the product of the marginals:

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

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,

$$E_{P(s(x,\mu_h)|Y=g)}[s(X,\mu_h)] = \frac{\sum_{z \in \mathcal{N}_g(x)} s(z,\mu_h)}{k_g}, \qquad (2)$$

where $k_g = |\mathcal{N}_g(x)|$ and the normalization parameters $\{\gamma_{gh}\}$ guarantee that the class-conditional models are in fact pmfs. Thus, the standard formulation of the local SDA classifier is:

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

This exponential SDA model admits two interpretations: a) it is the maximum likelihood exponential model with independent marginals, and b) it is the maximum entropy distribution that satisfies the set of mean constraints given by (2). The local SDA classifier possesses both the interpretability of generative models and the low bias of local classifiers, and has been shown to be competitive with state-of-the-art similarity-based classifiers [1].

The standard local SDA classifier is also consistent, in the sense that its classification error converges to the Bayes error [3], but the convergence may be hindered when degenerate local pmfs arise. Such is the case when the local mean constraint takes the form

$$E_{P(s(x,\mu_h)|Y=g)}[s(X,\mu_h)] = c, \; c \in \{\inf(\Omega), \sup(\Omega)\},$$

which can occur for example with small neighborhood sizes. When the mean constraints takes on an extremal value, it can only be satisfied by a Kronecker delta pmf: $P(s(X,\mu_h)|Y = g) = \delta(s(x,\mu_h) - c)$, which gives rise to the degenerate class-conditional joint $\prod_h P(s(X,\mu_h)|Y = g) \equiv 0$, and causes classification errors.

Previous work addressed this risk of degeneracy by either reverting to a local centroid classifier when degeneracy was detected [3], or by regularization of the estimated class-conditional pmfs [4]. Of a number of regularization approaches considered, the theoretical and empirical evidence favored regularizing the pmfs by forming a convex combination of the

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

local estimated SDA pmf with the average of the estimated local pmfs from the training samples:

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

where $P_{ave}(s_h|g)$ denotes the average of n marginal class-conditional local pmfs computed from the training set and the regularization parameter $\zeta\in[0,1]$. Analogously, one could regularize the joint class-conditional pmfs with the quantity $\prod_{h=1}^{G}P_{ave}(s_h|g)$. Both regularization methods produce smoother, non-exponential estimates of the local pmfs, thus enlarging the set of allowable class-conditional pmf models beyond the discrete exponential functions.

The first contribution of this paper is a Bayesian approach to estimating the local class-conditional pmfs based on a mutinomial model of the similarity distribution and a Dirichlet prior. Instead of regularizing the pmfs by convex combination, the Bayesian approach estimates the local similarity pmf as an expected posterior distribution, which has a closed-form solution.

The second contribution of this paper formulates the local SDA classifier in terms of the given pairwise similarities between the test sample x and all of its neighbors, rather than in terms of local centroids statistic.

A third constribution of this paper is a combined pairwise Bayesian local SDA classifier, which combined the Bayesian and pairwise similarity approaches. In the next sections, we discuss these contributions in detail.

III. Bayesian Local SDA

Here we detail how to implement SDA using a Dirichlet prior for the local pmf's, rather than assuming an exponential model.

A. The Bayesian Setup

For notational convenience, denote the local SDA pmf of interest as θ_{gh}, which is a pmf over the domain of similarities Ω such that $\theta_{gh}\in[0,1]^r$ where $r=|\Omega|$, and $\sum_{j=1}^{r}\theta_{gh}[j]=1$. Let Θ_{gh} be a random pmf whose support is the r-dimensional probability simplex, and which has realization θ_{gh}.

The Bayesian approach is to estimate the local SDA pmf $\hat{\theta}_{gh}$ as the *expected* pmf

$$\hat{\theta}_{gh}=E_{\Theta_{gh}}[\Theta_{gh}],$$

where the expectation of Θ_{gh} is taken with respect to its posterior probability distribution. By Bayes' rule, the posterior can be decomposed into the product of the likelihood and the prior divided by the probability of the data:

$$p(\theta_{gh}|D_{gh})=p(D_{gh}|\theta_{gh})p(\theta_{gh})/p(D_{gh}),\qquad(3)$$

where $p(\theta_{gh})$ is the prior probability of the pmf θ_{gh} and $p(D_{gh}|\theta_{gh})$ is the likelihood of the observed data under the assumption that the similarities are distributed according to θ_{gh}. In the standard local SDA model, the data over which the likelihood is evaluated are defined as the set of similarities between the neighbors of test point x from class g and the class h centroid μ_h: $D_{gh}=\{s(z_i,\mu_h)|z_i\in\mathcal{N}_g(x)\}$. The term $p(D_{gh})$ is the prior probability of the data, a normalizing factor. Then, the expected pmf is:

$$\begin{aligned}\hat{\theta}_{gh}&=&E_{\Theta_{gh}}[\Theta_{gh}]\\&=&\int_{\theta_{gh}}\theta_{gh}p(D_{gh}|\theta_{gh})p(\theta_{gh})/p(D_{gh})d\theta_{gh}.\end{aligned}\qquad(4)$$

To evaluate (4) with a closed-form solution, one would like the prior $p(\theta_{gh})$ to be a conjugate prior to the likelihood. For the standard local SDA discrete exponential model, the likelihood is a product of discrete exponentials functions over the finite, countable set Ω, and the exponent $\lambda_{gh}\in\mathbb{R}$: for this atypical exponential function no conjugate prior is known. To overcome this difficulty, one could give up the convenience of a closed-form solution and numerically evaluate (4) after choosing an appropriate, known pmf as the prior – for example based on expert knowledge of the pairwise similarities.

Numerically evaluating (4) has several drawbacks, first and foremost the need for numerical evaluation – for example by Monte Carlo methods – which is time consuming. Also, the lack of a closed-form solution makes it difficult to characterize the properties of the expected pmf, and more generally hinders the conceptual comparison to other classifiers.

Here we propose the simpler approach of expanding the family of allowable local similarity pmfs beyond discrete exponentials to be the entire family of multinomial distributions. The defining assumption of the proposed Bayesian local SDA approach is that the local similarities being modeled are multinomially distributed. For the multinomial distribution, the conjugate prior is the Dirichlet distribution [13], and the expected pmf (4) has an intuitive, closed-form solution: it is the expected value of a Dirichlet-distribution posterior.

B. The Dirichlet-Multinomial Similarity Model

Given a sample z assumed to be from class g and a neighborhood $\mathcal{N}(z)$, assume that its similarity to the local centroid $\mu_h\in\mathcal{N}_h(z)$ is distributed according to the pmf θ_{gh}. Suppose there are $k_g=|\mathcal{N}_g|$ such observations from the same neighborhood, and let the vector m be the histogram of occurrences of the jth similarity value $\omega_j\in\Omega$. That is, $m[j]=\sum_{i=1}^{k_g}\delta(s(z_i,\mu_h)-\omega_j)$, for $z_i\in\mathcal{N}_g$, where δ is the Kronecker delta function. Assuming these locally observed similarity values are independently drawn, the likelihood of a particular histogram m is

$$p(m)\sim\text{Mult}(\theta_{gh})=k_g!\prod_{j=1}^{r}\frac{1}{m[j]!}\left(\theta_{gh}[j]\right)^{m[j]}.$$

Following the Bayesian framework, we assume that θ_{gh} is a realization of a random pmf Θ_{gh}, which we assume is distributed as the Dirichlet distribution (which is a conjugate prior of the multinomial distribution):

$$p(\theta_{gh})\sim\text{Dir}(\alpha_{gh}),$$

where the Dirichlet vector parameter $\alpha_{gh}\in\mathbb{R}_+^r$ is the unnormalized mean pmf such that: $E[\theta_{gh}]=\alpha_{gh}/\sum_{j=1}^{r}\alpha_{gh}[j]$.

For the multinomial-Dirichlet conjugate pair, the posterior is the Dirichlet distribution

$$p(\theta_{gh}|m) \sim \text{Dir}(\alpha_{gh} + m).$$

Then, the expected posterior pmf $\hat{\theta}_{gh}$ in (4) is simply the expected value of the posterior Dirichlet distribution,

$$\hat{\theta}_{gh} = \frac{\alpha_{gh} + m}{\sum_{j=1}^{r} \alpha_{gh}[j] + m[j]}. \tag{5}$$

The resulting *Bayesian local SDA* classifier classifies a test sample x as the class $\hat{y}$ that solves

$$\arg \min_{f=1,\ldots,G} \sum_{g=1}^{G} C(f,g) \left(\prod_{h=1}^{G} \frac{\alpha_{gh} + m}{\sum_{j=1}^{r} \alpha_{gh}[j] + m[j]} \right) P(g). \tag{6}$$

We estimate each Dirichlet prior parameter $\{\alpha_{gh}\}$ by maximizing the likelihood (ML) of the observed histograms of local similarities, $D_{gh} = \{m^i\}$, $i = 1, \ldots n$,

$$\alpha_{gh}^{*} = \arg \max_{\alpha_{gh}} p(D_{gh}|\alpha_{gh}).$$

To estimate α_{gh}^{*}, in this work we adopt the iterative numerical procedure described by Minka, which is based on the observation that the likelihood is distributed according to a multivariate Polya distribution [14].

Expression (5) provides an intuitive interpretation for the estimated marginal pmfs. For a given test point x assumed from class g, the probability of observing the similarity $s(x, \mu_h) = \omega_j$ to the centroid of class h is given by the a priori mean probability $\alpha_{gh}[j]$ updated with the histogram $m[j]$ of the occurrences of ω_j in x's local neighborhood $\mathcal{N}(x)$.

IV. Pairwise Local SDA

In this section we propose a variant of local SDA that models the probability of seeing all the given pairwise similarities in the neighborhood of the test sample, rather than only the similarities to the local centroids.

Given a test sample x assumed from class g and its k-sized neighborhood $\mathcal{N}(x)$, let its local similarity statistics be $\mathcal{T}(x) = \{s(x,z)|z \in \mathcal{N}(x)\}$, the set of the similarities between x and all its neighbors, and let the set of the similarities between x and its neighbors only from class h be denoted by $\mathcal{T}_h(x) = \{s(x,z)|z \in \mathcal{N}_h(x)\}$, so that $\mathcal{T}(x) = \bigcup_h \mathcal{T}_h(x)$. The defining property of pairwise local SDA is that the marginal probability of x is defined as the average of the estimated probabilities of the similarities of x to its neighbors from class h

$$\begin{aligned} P_h(\mathcal{T}_h(x)|Y=g) &\triangleq \frac{1}{k_h} \sum_{z \in \mathcal{N}_h(x)} \hat{P}(s(x,z)|Y=g) \\ &= \frac{1}{k_h} \sum_{z \in \mathcal{N}_h(x)} \gamma_{gh} e^{\lambda_{gh} s(x,z)}. \end{aligned} \tag{7}$$

The parameters $\{\lambda_{gh}\}$ are found by numerical optimization under the method-of-moment constraint that reflects the contribution of all the pairwise similarities between the neighbors of x: the expected similarity of a random sample to its neighbors from class h must be equal to the empirical average of the local similarities,

$$E_{P_h(\mathcal{T}_h(x)|Y=g)}[s(X,z)] = \frac{\sum_{z_i \in \mathcal{N}_g(x)} \sum_{z_j \in \mathcal{N}_h(x)} s(z_i, z_j)}{k_g k_h}. \tag{8}$$

Then the class-conditional pmfs are modeled as the product of the marginal class-conditional pmfs (making an independence assumption),

$$P(\mathcal{T}(x)|Y=g) = \prod_{h=1}^{G} P_h(\mathcal{T}_h(x)|Y=g), \tag{9}$$

and the test sample x is classified according to (1).

In the experiments we also consider the pairwise local SDA classifier where the marginal pmfs are estimated using the Bayesian approach described in Section III. For this combined pairwise Bayesian local SDA classifier, the local SDA models are estimated using the Dirichlet-multinomial Bayesian formulation based on all the pairwise local similarities. The classification rule is identical to (6), but the histogram m is the count of all pariwise similarities, that is $m[j] = \sum_{i=1}^{k_g} \sum_{l=1}^{k_h} \delta(s(z_i, z_l) = \omega_j)$ for $z_i \in \mathcal{N}_g(x), z_l \in \mathcal{N}_h(x)$.

V. Experiments

We present results for the proposed Bayesian local SDA and pairwise local SDA, and for the combined pairwise Bayesian local SDA classifier. We compare the proposed classifiers to the standard and regularized local SDA classifiers, and to the standard k-nearest neighbor classifier (in similarity space), and four state-of-the-art local SVM classifiers.

A. Data and Setup

We show results on experiments for five different similarity-based datasets from a variety of fields. These datasets and the classification software are available on the Web. [2]

The Amazon problem is to classify books as fiction or non-fiction, where the similarity between two books is the symmetrized percentage of customers who bought the first book after viewing the second 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 of metrics 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. The negative eigenfraction (sum of the magnitudes of the negative eigenvalues divided by the sum of all magnitudes) for the Amazon pairwise similarity matrix is 0.6%.

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 independent listeners, resulting in a perceptual

[2] http://staff.washington.edu/lucagc and http://idl.ee.washington.edu/similaritylearning

similarity from 2 to 10 [15]. This dataset is interesting because perceptual similarities are often non-metric, in that they do not satisfy the triangle inequality. The negative eigenfraction for Aural Sonar is 21%.

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 [16]. The self-similarity is set to 1. Like the Amazon dataset, this is a sparse dataset and most of the similarities equal to zero, with a large 40% negative eigenfraction.

The Protein problem is to classify 213 proteins into one of four protein classes based on a sequence-alignment similarity [17]. This problem is very well-suited to treating similarities as features because many of the first class proteins are consistently more similar to proteins from the second class [1]. The negative eigenfraction for Protein is 20%.

The Voting problem is to classify 435 representatives into two political parties based on their votes [18]. 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 [19]. The voting similarity is almost metric, with a very small eigenfraction of 0.06%.

The five 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 the tables. Cross-validation parameter sets were based on recommendations in previously published papers and popular usage. For all local classifiers, the choice of neighborhood sizes was $k \in \{2, 4, 8, 16, 32, 64, \min(n, 128)\}$ and voting ties are decided according to the prior for KNN. For regularized local SDA, the choices for the convex regularizing parameter were $\zeta \in \{10^{-6}, 10^{-3}, 0.01, 0.1, 0.5, 0.9\}$. For all SVMs, the standard C parameter choices were $10^{-3}, 10^{-2}, \ldots, 10^{5}$. Both ζ and C are cross-validated in nested loops with k. Multi-class implementations of the SVM classifiers used $\binom{n}{2}$ pairwise classifiers.

B. Classification Results

Results are shown in Table I and Table II, 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 ($p = 0.05$).

Table I compares the proposed local SDA extensions to previously-published local SDA variants, and Table II shows the performance of the combined pairwise Bayesian local SDA compared to state-of-the-art variants of local SVM similarity classifiers based on KNN-SVM [1]. The most interesting result is that modeling the local pairwise similarities rather than the local centroid similarities results in a dramatic decrease in error for the Protein dataset from 19.42% to 9.77%. Combining the pairwise model with the Bayesian estimation further reduces the error to 1.63% on the Protein dataset - which is statistically tied with the best local SVM variant, as shown in Table II. Apart from the excellent performance on the Protein dataset, the pairwise Bayesian local SDA is statistically tied with the other local SDA variants.

Combined with the standard local centroid model, the Bayesian local SDA (top row of Table I) is not seen to provide a statistically significant improvement over local SDA. We believe this dependence on the type of similarity statistic being modeled emerges because the similarity to local centroids is more likely to be monotonic over the similarity domain Ω. The exponential model of local SDA better captures monotonic probability, whereas the multinomial model used in the Bayesian approach is too flexible, and thus induces too much estimation variance. In contrast, when modeling pairwise similarities, the true probability distributions of the similarities are more diverse, and are more likely to be unimodal than monotonic, and thus the multinomial model's flexibility becomes an advantage.

VI. Discussion and Hypotheses

We proposed a Bayesian model for estimating class-conditional distributions of similarities to reduce the chance of model degeneracy, and proposed modeling local pairwise similarities to reduce the variance due to the randomness of the local centroid. The experimental evidence shows that the resulting increased flexibility in the local SDA classifier makes it much more competitive with state-of-the-art local SVM methods for similarity-based learning, but with the added advantage of being a generative classifier, and thus providing natively probabilistic outputs and greater interpetability.

We hypothesize that even better results are possible with generative similarity-based classifiers. The SVM approaches require forming a PSD kernel from the similarities, and although in some cases that may provide useful regularization, in other cases that may cause the loss of useful information for the learning problem. The SDA approach models the similarities directly and does not require such restrictions or approximations, and thus for very indefinite similarities may have a performance advantage over kernel-based approaches.

The computational load for all SDA-type of classifiers is comparable. Just like local SDA, all other variants require estimating G^2 local pmfs for each test sample, which we accomplish quickly using standard function minimization and maximum likelihood methods. The Bayesian, pairwise, and regularized local SDA classifiers additionally require iterating through the entire training set to form n sets of local pmfs needed to estimate the Dirichlet model parameters and the regularizing pmfs, respectively, for each cross-validated neighborhood size. For large datasets, this procedure can be time-consuming, but it need be done only once, thus does not preclude using the proposed methods for real-time similarity-based applications. More discussion on computational issues was previously published [4].

We adopted the multinomial-Dirichlet conjugate pair for Bayesian local SDA. Other choices for the similarity pmfs and the prior can be considered, with the goal of simplifying the model parameter estimation for the posteriors. However,

TABLE I

AVERAGE (STANDARD DEVIATION) OF THE PERCENT TEST ERROR OVER 20 RANDOM TEST/TRAIN SPLITS FOR THE SDA-TYPE CLASSIFIERS. 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)	Protein (4 classes)	Voting (2 classes)
Bayesian local SDA	**10.79 (8.79)**	19.75 (8.50)	**11.56 (4.71)**	22.91 (4.85)	**6.84 (1.95)**
pairwise local SDA	**12.37 (9.40)**	**16.25 (8.09)**	**11.56 (4.71)**	9.77 (5.97)	**7.07 (2.51)**
pairwise Bayesian local SDA	**11.58 (8.64)**	**16.25 (8.25)**	**11.56 (4.71)**	**1.63 (2.73)**	**6.90 (2.01)**
Local SDA	**12.37 (9.40)**	18.75 (7.05)	**11.56 (4.71)**	19.42 (6.84)	**6.90 (2.33)**
Regularized local SDA (marginal)	**12.11 (9.52)**	**18.75 (8.25)**	**11.56 (4.71)**	20.00 (7.64)	**6.72 (2.25)**
Regularized local SDA (joint)	**11.58 (9.60)**	**19.25 (8.47)**	**11.56 (4.71)**	19.65 (7.06)	**6.78 (2.33)**

TABLE II

AVERAGE (STANDARD DEVIATION) OF THE PERCENT TEST ERROR OVER 20 RANDOM TEST/TRAIN SPLITS, FOR THE PAIRWISE BAYESIAN LOCAL SDA CLASSIFIER AND FIVE OTHER CLASSIFIERS. 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)	Protein (4 classes)	Voting (2 classes)
pairwise Bayesian local SDA	11.58 (8.64)	16.25 (8.25)	**11.56 (4.71)**	**1.63 (2.73)**	6.90 (2.01)
KNN (on similarities)	12.11 (9.21)	15.75 (5.68)	19.48 (5.20)	30.00 (9.63)	5.69 (1.95)
KNN-SVM (clip)	**7.37 (6.48)**	**13.75 (7.59)**	13.75 (5.17)	11.86 (5.64)	**5.11 (2.40)**
KNN-SVM (flip)	**7.63 (6.27)**	**14.00 (5.98)**	13.02 (4.96)	**1.74 (2.49)**	**5.11 (2.37)**
KNN-SVM (shift)	**7.63 (6.27)**	**14.00 (7.00)**	13.33 (5.03)	30.23 (8.80)	**5.23 (2.43)**
KNN-SVM (sims-as-features)	13.68 (8.08)	**13.00 (6.16)**	14.58 (5.53)	29.65 (10.18)	**5.40 (1.40)**

the potential benefit of simpler computations must be considered jointly with the interpretability and relevance of the chosen models to the similarity-based classification problem. For instance, a negative Dirichlet prior [20] would eliminate the need for ML estimation of α_{gh}, but would be difficult to interpret and could introduce too much model bias by enforcing unrealistic sparseness in the estimated pmfs. On the contrary, in practice, similarity pmfs are often smooth. Research into suitable models for Bayesian SDA is ongoing.

Future extensions might incorporate smoothness and monotonicity constraints into the multinomial-Dirichlet model. The multinomial distribution is particularly well-suited for categorical random variables, for which the pmfs can be multimodal. However, similarities are ordinal, and the similarity pmf is generally monotonic. Lastly, we hypothesize that adopting the pairwise similarity approach rather than the single class centroid approach will provide the most flexibility while retaining good performance.

ACKNOWLEDGMENT

This work was supported by the United States Office of Naval 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, M. R. Gupta, and A. J. Koppal, “Generative models for similarity-based classification,” *Pattern Recognition*, vol. 41, no. 7, pp. 2289–2297, July 2008.

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

[4] ——, “Regularizing the local similarity discriminant analysis classifier,” in *Proc. 8th Intl. Conf. Machine Learning and Applications*, December 2009.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] H. Zhang, A. C. Berg, M. Maire, and J. Malik, “SVM-KNN: discriminative nearest neighbor classification for visual category recognition,” *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2126 – 2136, 2006.

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

[13] T. S. Ferguson, “A Bayesian analysis of some nonparametric problems,” *Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[14] T. P. Minka, “Estimating the Dirichlet distribution,” February 2003, available at `http://research.microsoft.com/m̃inka`.

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

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

[17] T. Hofmann and J. Buhmann, “Pairwise data clustering by deterministic annealing,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 1, January 1997.

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

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

[20] M. Bicego, M. Cristani, and V. Murino, “Sparseness achievement in hidden Markov models,” *Proc. IEEE Int. Conf. on Image Analysis and Processing (ICIAP07)*, pp. 67–72, 2007.