Maximum Entropy Generative Models for Similarity-based Learning

Maya R. Gupta
Dept. of EE
University of Washington
Seattle, WA 98195
gupta@ee.washington.edu

Luca Cazzanti
Applied Physics Lab
University of Washington
Seattle, WA 98195
luca@apl.washington.edu

Anjali J. Koppal
Dept. of EE and CS
University of California
Berkeley, CA
ajkoppal@berkeley.edu

***Abstract*— A generative model for similarity-based classification is proposed using maximum entropy estimation. First, a descriptive set of similarity statistics is assumed to be sufficient for classification. Then the class conditional distributions of these descriptive statistics are estimated as the maximum entropy distributions subject to empirical moment constraints. The resulting exponential class conditional distributions are used in a maximum a posteriori decision rule, forming the *similarity discriminant analysis* (SDA) classifier. The relationship between SDA and the quadratic discriminant analysis classifier is discussed. An example SDA classifier is given that uses the class centroids as the descriptive statistics. Compared to the nearest-centroid classifier, which is also based only on the class centroids, simulation and experimental results show SDA consistently improves performance.**

I. Similarity-based Classification

In similarity-based classification, the problem is to classify a test sample x given only the pairwise similarities between x and a set of training samples $\{x_i\}$, $i = 1, \ldots, n$, and given the pairwise similarity between any two of the training samples [1]–[4]. The training samples' class labels are given, and are denoted $\{y_i\}$. A similarity function s is a mapping that accepts two samples x, z from some sample space $x, z \in \mathcal{B}$, and returns a real number. That is, $s : \mathcal{B} \times \mathcal{B} \to \Omega$, where $\Omega \subset \mathbb{R}$. It is useful to think of the sample space $\mathcal{B}$ as an abstract space, such as "the space of all proteins," or "the space of all blogs." The similarity $s(x, z)$ is some judgement of how near samples x and z are, but similarities are not required to satisfy metric properties or any specific mathematical properties. The term "similarity-based classification" is also used when the given information is "dissimilarities," where a dissimilarity is a judgement of how far two samples are, but is not required to satisfy any specific mathematical properties.

Similarity-based learning is a flexible learning paradigm. Similarity-based learning is a useful approach when samples are described by categorical variables. For example, DNA is described as a sequence of unordered bases, A, T, G, and G. Similarity-based learning is of course appropriate when the similarity or dissimilarity between samples is not a metric. For example, driving-times between any two given locations is not a metric, as it is often not symmetric and can violate the triangle inequality. Categorical variables and non-metric similarities/disssimilarities are common in fields such as bioinformatics, information retrieval, and natural language processing [1], [4]. Also, similarity-based learning may be a better model than standard Euclidean-space learning for how humans classify, as psychologists have shown that metrics do not account for human judgements of similarity in complex situations [5]–[7]. Laub et al. have shown that nonmetric similarities lead to information that can be useful for pattern recognition [8].

The simplest method for similarity-based classification is the nearest neighbor classifier, which determines the most similar training sample to the test sample, and classifies the test sample as its most-similar neighbor's class. In fact, nearest neighbor classifiers using a tangent distortion [9] and a shape similarity metric [10] have both been shown to achieve very low error on the MNIST character recognition task.

Similarity-based classifiers that model a generating distribution have not been previously proposed. A simple model-based approach is the nearest centroid classifier [11]. The nearest-centroid classifier finds a centroid μ_h for the hth class:

$$\mu_h = \arg\max_{\mu \in \mathcal{X}_h} \sum_{z \in \mathcal{X}_h} s(z, \mu), \tag{1}$$

where $\mathcal{X}_h$ is the set of training samples from class h. Then, a test sample x is classified as the class $\hat{y}$, where

$$\hat{y} = \arg\max_h s(x, \mu_h) \tag{2}$$

The nearest-centroid classifier is analogous to the nearest-mean classifier if samples are described as Euclidean feature vectors, and each class is characterized by its mean Euclidean feature vector.

In this paper, we propose a generative model for similarity-based classification, which we term similarity discriminant analysis (SDA). Although the generative architecture is quite general, we consider in more depth SDA using class centroids, and show that compared to the nearest-centroid model classifier and to nearest-neighbor classification, the log-linear SDA classifier can achieve better results in simulation and on real data.

II. Generative Architecture for Similarity-Based Classification

Let a test sample x be a realization of the random variable $X \in \mathcal{B}$, where $\mathcal{B}$ is the sample space. Let the similarity function be some function $s : \mathcal{B} \times \mathcal{B} \rightarrow \Omega$, where $\Omega \subset \mathbb{R}$. In this paper, we assume that the sample space $\mathcal{B}$ is finite and discrete, such that the space of the possible pairwise similarities Ω is also finite and discrete. Let $Y \in \mathcal{G}$ be the class label associated with x, where $\mathcal{G}$ is a finite set of classes. Let $C(g, h)$ be the cost of classifying x as class g if the true class is h.

An optimal classifier is the theoretical Bayes classifier [12], which assigns a test sample x the class $\hat{y}$ that minimizes the expected misclassification cost,

$$\hat{y} = \arg \min_{f=1,\ldots,G} \sum_{g=1}^{G} C(f, g) P(Y = g|x), \tag{3}$$

where $C(f, g)$ is the cost of classifying the test sample x as class f if the true class is g. In practice the distribution $P(g|x)$ is generally unknown.

For the generative model, assume that the relevant information about X's class label is captured by some finite set $\mathcal{T}(X)$ descriptive statistics. For example,

$$\mathcal{T}(X) = \{s(X, \mu_1), s(X, \mu_2), \ldots, s(X, \mu_G)\}. \tag{4}$$

Under this assumption the classification rule (3) for a particular test sample x is to classify x as class the $\hat{y}$ that solves

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

Using Bayes rule, this is equivalent to the problem

$$\arg \min_{f=1,\ldots,G} \sum_{g=1}^{G} C(f, g) P(\mathcal{T}(x)|Y = g) P(Y = g). \tag{5}$$

Next, we assume that each unknown class conditional distribution $P(\mathcal{T}(x)|Y = g)$ has the same average value as the training sample data from class g. That is, we assume that the mth descriptive statistic $\mathcal{T}_m(x)$ has mean equal to the training sample mean:

$$E_{P(\mathcal{T}(x)|g)}[\mathcal{T}_m(X)] = \frac{1}{n_g} \sum_{z \in \mathcal{X}_g} \mathcal{T}_m(z), \tag{6}$$

for $g = 1, \ldots, G$ and $m = 1, \ldots, M$. Given these $M \times G$ constraints, there is some compact and convex feasible set of G class conditional distributions $P(\mathcal{T}(x)|Y = g)$. A feasible solution will always exist because the constraints are based on the data.

As prescribed by Jaynes' principle of maximum entropy [13], we propose selecting the unique class conditional distributions that satisfy (6) and maximize entropy. Maximum entropy distributions have the maximum possible uncertainty, and in that sense are the least assumptive solution. Given a set of moment constraints, the maximum entropy solution is known to have exponential form. Selecting the maximum entropy distribution subject to constraints is analogous to the generative classifier *quadratic discriminant analysis* (QDA). QDA models each class conditional distribution as a Gaussian [12], which is the maximum entropy distribution if the distribution's mean and covariance are constrained to match the sample mean and covariance for each class.

For the gth class, solving the M constraints specified by (6) for the maximum entropy distribution yields

$$\hat{P}(\mathcal{T}(x)|g) = \prod_{m=1}^{M} \gamma_{gm} e^{\lambda_{gm} \mathcal{T}_m(x)} \tag{7}$$

$$= \prod_{m=1}^{M} \hat{P}(\mathcal{T}_m(x)|g), \tag{8}$$

where the parameters $\{\lambda_{gm}, \gamma_{gm}\}$ have unique solutions which satisfy the constraints defined by (6). The equality given in (8) shows that under the maximum entropy assumption the statistics comprising the set $\mathcal{T}(x)$ are conditionally independent given the class label. Thus, one could equivalently describe this model as the maximum entropy solution given the constraints

$$E_{P(\mathcal{T}_m(x)|g)}[\mathcal{T}_m(X)] = \frac{1}{n_g} \sum_{z \in \mathcal{X}_g} \mathcal{T}_m(z) \tag{9}$$

for $g = 1, \ldots, G$ and $m = 1, \ldots, M$, because the estimated $P(\mathcal{T}(x)|Y = h)$ is the same.

Substituting the maximum entropy solution (7) into (5) creates the SDA classification rule: classify x as the class $\hat{y}$ which solves

$$\arg \max_{f=1,\ldots,G} \sum_{g=1}^{G} C(f, g) P(g) \prod_{m=1}^{M} \gamma_{gm} e^{\lambda_{gm} T_m(x)} . \tag{10}$$

III. SDA Based on Class Centroids

We further investigate the two-class SDA classifier using the descriptive statistics given in (4) and zero-one misclassification costs (that is, $C(f, g) = 0$ if $f = g$ and $C(f, g) = 1$ otherwise). In this case, the SDA classification rule (10) becomes: choose class 1 if

$$\frac{\hat{P}(s(x, \mu_1)|Y = 1)}{\hat{P}(s(x, \mu_1)|Y = 2)} \frac{\hat{P}(s(x, \mu_2)|Y = 1)}{\hat{P}(s(x, \mu_2)|Y = 2)} \frac{P(Y = 1)}{P(Y = 2)} > 1. \tag{11}$$

Applying the maximum entropy solution for the class conditional distributions, (11) becomes: choose class one if

$$\frac{\gamma_{11} e^{\lambda_{11} s(x, \mu_1)}}{\gamma_{21} e^{\lambda_{21} s(x, \mu_1)}} \frac{\gamma_{12} e^{\lambda_{12} s(x, \mu_2)}}{\gamma_{22} e^{\lambda_{22} s(x, \mu_2)}} \frac{P(Y = 1)}{P(Y = 2)} > 1. \tag{12}$$

This SDA classifier uses the same information about the test sample as the nearest-centroid classifier, $s(x, \mu_1)$ and $s(x, \mu_2)$, but models the probability distribution of these statistics under the hypothesis that the sample belongs to class one or to class two. The probability distributions of the similarities capture the characteristic average deviation for each class and the average cross-class deviations. It is helpful to group the terms in (11) into the ratio term $\hat{P}(s(x, \mu_1)|Y = 1)/\hat{P}(s(x, \mu_1)|Y = 2)$

and the ratio term $\hat{P}(s(x,\mu_2)|Y=1)/\hat{P}(s(x,\mu_2)|Y=2)$. The first of these ratio terms establishes whether the similarity between the test sample x and the class one centroid μ_1 is better explained probabilistically by assuming x is from class one or from class two. Likewise, the second ratio term establishes whether the similarity $s(x,\mu_2)$ is better explained probabilistically by the hypothesis that x is from class one or from class two.

The SDA classifier given in (12) uses the class centroids, but is more flexible than the nearest centroid classifier. For example, consider the case in which class one training samples are tightly clustered around μ_1, but class two training samples have on average low similarity to μ_2. Then even if a test sample x is slightly more similar to μ_1 such that $s(x,\mu_1) > s(x,\mu_2)$, SDA can learn that class one points should be very similar to μ_1, and can correctly classify x as a class two sample. This is analogous to the action of quadratic discriminant analysis in the case that class one's variance is very low compared to class two's variance.

It is also helpful to consider the effect of the exponential form of the class conditional distributions. Assume for this analysis that the possible similarities are uniformly spaced between 0 and 1. Then if the average empirical similarity between class one training samples and the class one centroid μ_1 is greater than .5, the exponential distribution $\hat{P}(s(x,\mu_1)|Y=1)$ will necessarily be an increasing function of similarity. Thus, high probability will be given to test samples that have $s(x,\mu_1) > .5$. However, if the average empirical similarity between class one training samples and the class one centroid μ_1 is less than .5, the exponential distribution $\hat{P}(s(x,\mu_1)|Y=1)$ will be a decreasing function, and low probability will be given to test samples that have $s(x,\mu_1) > .5$. Thus, if class one training samples are all relatively dissimilar to the class one centroid, then a test sample that is too similar to the class one centroid is assigned a low probability of being a class one sample.

A simpler method to generalize the nearest centroid classifier to take into account the different distributions of each class would be to directly take into account the average similarity $\bar{s}_{gg}$ between class g training samples and a class g centroid μ_g, where

$$\bar{s}_{gg} = (1/n_g) \sum_{x_j \in \mathcal{X}_g} s(x_j, \mu_g).$$

Then, classify a test sample x as class $\hat{y}$ where

$$\hat{y} = \arg\max_g \frac{s(x,\mu_g)}{\bar{s}_{gg}}. \tag{13}$$

This is analogous to the Gaussian-derived rule of classifying by the distances to the class means inversely weighted by each class's standard deviation: $\|x-\mu_g\|/\sigma_g$. We term the classifier given in (13) the *nearest centroid adjusted* classifier, and include it in our experiments detailed in the next section.

IV. Experiments

Experiments were done to compare SDA using the centroidal descriptive statistics given in (4) to nearest centroid, nearest centroid adjusted (13), and nearest neighbors for similarity-based classification.

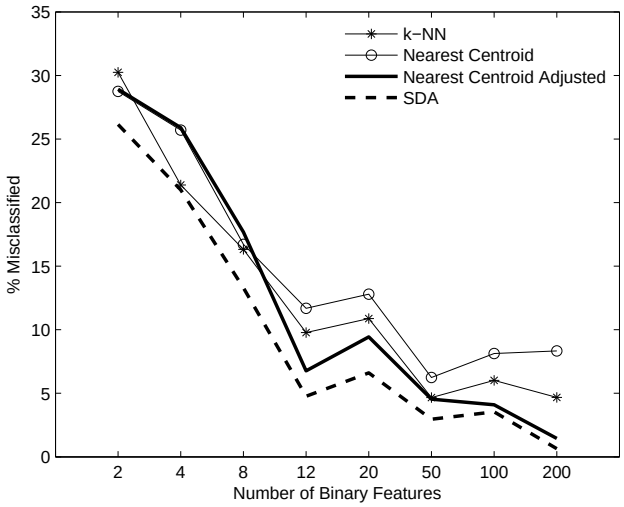

Fig. 1. Results for Perturbed Centroids Simulation.

A. Perturbed Centroids Simulation

We begin with a simple simulation where each class is generated by perturbing one centroidal sample; thus the nearest-centroid classifiers are a good model for this simulation. There are two classes, and each class is defined by one prototypical set of n binary features, c_1 or c_2, where c_1 and c_2 are each drawn uniformly and independently from $\{0,1\}^n$. A training or test sample z drawn from class g has the ith feature $z[i] = c_g[i]$ with probability $1-p_g$, and $z[i] \neq c_g[i]$ with perturbation probability p_g. Each time the simulation is run, the perturbation probabilities p_1, p_2 were drawn randomly from a uniform distribution, $p_1 \sim unif[0, \frac{1}{3}]$, and $p_2 \sim unif[0, 1/2]$. The two classes are equally likely. There were 20 training samples, and 1000 test samples per run, and the simulation was run 20 times.

For this simulation, we used the simple *counting similarity*. That is, the similarity between x and z is the number of features in which their feature value agrees. The results are reported in Figure 1, and show that SDA performs consistently better than the other classifiers.

B. Protein Data

Many bioinformatics prediction problems are formulated in terms of pairwise (dis)similarities. An example is the protein data set used in [4]. The dissimilarity values are calculated using a sequence alignment program, which measures the number of amino acids that differ between two sequences [14]. Following [4], we used the 213 proteins that had the class labels, "HA" (72 samples), "HB" (72 samples), "M" (39 samples) , and "G" (30 samples). Changing the similarity-based classifiers described in this paper from similarities to dissimilarities is straightforward: class centroids are chosen to maximize sum-dissimilarity rather than to minimize

TABLE I
CLASSIFICATION RESULTS FOR FOUR PROTEIN PROBLEMS.

Classifier	% Misclassified for Each Problem			
	HA	HB	M	G
1 Nearest Neighbor	77%	51%	13%	13%
3 Nearest Neighbors	85%	55%	15 %	14%
5 Nearest Neighbors	81%	50%	16%	14%
Nearest Centroid	30%	42%	**0%**	12%
Nearest Centroid Adjusted	30%	**25%**	4%	22%
SDA	**29%**	29%	**0%**	**1%**

sum-similarity, the nearest-neighbor minimizes dissimilarity, and the SDA model uses the descriptive statistics $\mathcal{T}(x) = \{d(x,\mu_1), d(x,\mu_2)\}$.

For SDA, the class priors were estimated to be the empirical probability of seeing a sample from each class, with Laplace correction. The set of possible dissimilarities Ω is needed to solve for the SDA parameters λ and γ, but was not directly available, so Ω was approximated as the set of empirical similarities that occurred in the training samples' similarity matrix. The class centroids were as defined in (1).

Table I shows the percentage misclassification for the four one-class-vs.-the-rest problems for each of the classifiers, calculated as the leave-one-out error on the 213 samples (the results were rounded for display). For this problem the nearest centroid, nearest centroid adjusted, and SDA were implemented as mixtures where each of the four classes was represented by its own centroid (and in the case of SDA, its own class conditional distribution). SDA performs better than the nearest centroid model classifiers and nearest neighbor classifiers. The relatively low error rates on class "M" and class "G" suggest that those classes are well-modeled by SDA's centroidal model, but that classes "HA" and "HB" are not.

V. RELATIONSHIP OF SDA TO OTHER CLASSIFIERS

A. Relationship of SDA to QDA

Quadratic discriminant analysis (QDA) is a generative classifier for standard metric-based learning (but not similarity-based learning). QDA models each class by a Gaussian distribution, where the mean $\hat{\mu} \in \mathbb{R}^d$ and covariance matrix $\hat{\Sigma} \in \mathbb{R}^d \times \mathbb{R}^d$ are estimated from the training samples for each class, often using maximum likelihood. Then, a test point $x \in \mathbb{R}^d$ is classified by determining which class conditional Gaussian generating distribution is most likely to have generated the test point. The Gaussian distribution is the maximum entropy distribution given that the generating distribution is constrained to have mean $\hat{\mu}$ and covariance matrix $\hat{\Sigma}$.

If one considers $z = \mathcal{T}(x)$ to be a feature vector of the test sample x, and $z_i = \mathcal{T}(x_i)$ to be a feature vector of the ith training sample x_i, then SDA is similar to QDA. In both SDA and QDA, a class conditional distribution is estimated to be the maximum entropy distribution given some empirical moment constraints. And then a decision is made based on the estimated class conditional likelihood of the feature vector $z = \mathcal{T}(x)$ in the case of SDA, or the Euclidean test feature vector in the case of QDA.

B. Relationship of SDA to Discriminant Analysis on Similarity Features

Similarity-based classification problems can be turned into standard Euclidean-based learning problems by taking the $n \times 1$ vector of similarities between a test sample and the n training samples, and using it as an n-dimensional Euclidean feature vector [2], [15], [16]. However, this approach turns the similarity-based classification into a standard metric statistical learning problem with n training samples in an n-dimensional feature space, with the concomitant *curse of dimensionality* difficulties [12]. Duin et al. propose dealing with the resulting curse of dimensionality problem by using a regularized linear discriminant analysis classifier on the n-dimensional feature space [2], [16]. We refer to this method as *discriminant analysis on similarity features*. Their results show that, on average over their different experiments, linear classifiers built on the similarity vectors achieve similar errors as the 1-nearest neighbor similarity-based classifier, except in cases of severe noise, where the 1-nearest neighbor has high error.

In this section, we further analyze *discriminant analysis on similarity features*, and compare it with SDA. We formulate the problem in terms of dissimilarities as done in [2], and for simplicity consider only the two-class problem where, without loss of generality, the classes are equally likely a priori. Let $D(x)$ be the $n \times 1$ feature vector for a test point x, such that the ith component of $D(x)$ is $d(x, x_i)$, for $i = 1, \ldots, n$ and where d is a dissimilarity function such that $d : \mathcal{B} \times \mathcal{B} \rightarrow \Omega$, $\Omega \subset \mathbb{R}$.

Discriminant analysis on similarity features classifies a test sample based on the discriminant

$$f(D(x)) = D(x)^T \Sigma^{-1}(\tilde{\mu}_1 - \tilde{\mu}_2) - .5(\tilde{\mu}_1 + \tilde{\mu}_2)^T \Sigma^{-1}(\tilde{\mu}_1 - \tilde{\mu}_2), \tag{14}$$

where the jth component of $\tilde{\mu}_1$ is

$$\frac{1}{n_1} \sum_{x_i \in \mathcal{X}_1} d(x_i, x_j), \tag{15}$$

$\tilde{\mu}_2$ is defined analogously, and Σ is the (unbiased) maximum likelihood estimate of the within-class pooled covariance, defined as:

$$\hat{\Sigma} = \frac{1}{(n-2)} \sum_{g=1}^{2} \sum_{x_i \in \mathcal{X}_g} (D(x_i) - \tilde{\mu}_g)(D(x_i) - \tilde{\mu}_g)^T \tag{16}$$

If there are n training samples, then there are $n \times n$ parameters to estimate for $\hat{\Sigma}$, which is generally ill-posed. Duin et al. suggest regularizing $\hat{\Sigma}$ by forming a convex combination between the empirical within-class pooled covariance and the identity matrix. To simplify the present analytic comparison, suppose that the true pooled covariance is the identity matrix I, and optimistically suppose that the pooled covariance is estimated perfectly, so that $\hat{\Sigma} = I$. Then the classification rule becomes: classify as class one if $f(D(x)) > 0$, that is, if

$$D(x)^T \tilde{\mu}_1 - 0.5 \tilde{\mu}_1^T \tilde{\mu}_1 > D(x)^T \tilde{\mu}_2 - 0.5 \tilde{\mu}_2^T \tilde{\mu}_2. \tag{17}$$

This decision rule is based on whether the feature vector of dissimilarities $D(x)$ is better correlated with $\tilde{\mu}_1$, the vector of mean distances to each training sample from class one training samples, or to $\tilde{\mu}_2$, the vector of mean distances to each training sample from class two training samples, where these correlations are offset by the self-correlations of $\tilde{\mu}_1$ and $\tilde{\mu}_2$.

Although in this paper we focused on SDA using the nearest centroid descriptive statistics given in (4), a closer comparison between SDA and *discriminant analysis on similarity features* is possible if the SDA classifier uses for descriptive statistics the set of dissimilarities to each training sample

$$\mathcal{T}(x) = \{d(x, x_1), d(x, x_2), \ldots, d(x, x_n)\}. \tag{18}$$

Then the SDA decision rule becomes: classify as class one if

$$\prod_{j=1}^{n} \gamma_{j1} e^{\lambda_{j1} d(x, x_j)} > \prod_{j=1}^{n} \gamma_{j2} e^{\lambda_{j2} d(x, x_j)}, \tag{19}$$

where λ_{jh} satisfies for $h \in \{1, 2\}$ the constraint specified in (9), which is:

$$\sum_{d(x,x_j)\in\Omega} d(x, x_j) \gamma_{jh} e^{\lambda_{jh} d(x, x_j)} = \frac{1}{n_h} \sum_{x_i \in \mathcal{X}_h} d(x_i, x_j). \tag{20}$$

Take the logarithm of (19),

$$\sum_{j=1}^{n} (\ln(\gamma_{j1}) + d(x, x_j)\lambda_{j1}) > \sum_{j=1}^{n} (\ln(\gamma_{j2}) + d(x, x_j)\lambda_{j2}),$$

or equivalently, where λ_1 is a vector with jth component λ_{j1},

$$\left(\sum_{j=1}^{n} \ln(\gamma_{j1}) \right) + D(x)^T \lambda_1 > \left(\sum_{j=1}^{n} \ln(\gamma_{j2}) \right) + D(x)^T \lambda_2. \tag{21}$$

Thus, using the descriptive statistics given in (18), the SDA rule given in (21) and the *discriminant analysis on similarity features* rule (17) (assuming the estimated covariance was the identity matrix) have the same form. Also, the constants in both rules are due to the normalization of the underlying probability model. However, in (17) the tested correlations are between the test feature vector $D(x)$ and each class centroid $\tilde{\mu}$, whereas in (21) the tested correlations are between the test feature vector $D(x)$ and the parameter vector λ.

VI. Discussion and Some Open Questions

In this paper, we have proposed a maximum-entropy based architecture for generative similarity-based classifiers, resulting in a log-linear classifier we term SDA. We have shown that the nearest-centroid SDA classifier can perform better than related nearest-centroid classifiers that use the same information for both a simple simulation and a real dataset, as well as performing better than nearest neighbor classification. An advantage of SDA over non-probabilistic models such as nearest centroid classifiers, is that it estimates probabilities. With estimated classifier probabilities, it is straightforward to account for class prior probabilities and different misclassification costs.

As with LDA and QDA, the power of a generative classifier depends on how well its model matches the true class conditional distributions. Gaussian mixture model classifiers are a flexible approach to Euclidean-based learning, and SDA using mixture models should have many of the same benefits. The design of general SDA mixture models is an open research question.

References

[1] M. Bicego, V. Murino, M. Pelillo, and A. Torsello, "Special issue on similarity-based classification," *Pattern Recognition*, vol. 39, October 2006.

[2] 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.

[3] D. W. Jacobs, D. Weinshall, and Y. Gdalyahu, "Classification with nonmetric distances: Image retrieval and class representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 6, pp. 583–600, June 2000.

[4] S. Hochreiter and K. Obermayer, "Support vector machines for dyadic data," *Neural Computation*, vol. 18, no. 6, pp. 1472–1510, 2006.

[5] A. Tversky, "Features of similarity," *Psychological Review*, no. 84, pp. 327–352, 1977.

[6] A. Tversky and I. Gati, "Studies of similarity," in *Cognition and Categorization*, E. Rosch and B. Lloyd, Eds. Hillsdale, N.J.: Earlbaum, 1978.

[7] I. Gati and A. Tversky, "Weighting common and distinctive features in perceptual and conceptual judgments," *Cognitive Psychology*, no. 16, pp. 341–370, 1984.

[8] 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.

[9] P. Simard, Y. L. Cun, and J. Denker, "Efficient pattern recognition using a new transformation distance," *Advances in Neural Information Processing Systems 5*, pp. 50–68, 1993.

[10] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, April 2002.

[11] D. Weinshall, D. W. Jacobs, and Y. Gdalyahu, "Classification in non-metric spaces," *Advances in Neural Information Processing Systems 11*, pp. 838–844, 1999.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer-Verlag, 2001.

[13] E. T. Jaynes, "On the rationale for maximum entropy methods," *Proc. of the IEEE*, vol. 70, no. 9, pp. 939–952, September 1982.

[14] 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.

[15] T. Graepel, R. Herbrich, and K. Obermayer, "Classification on pairwise proximity data," *Advances in Neural Information Processing Systems 11*, pp. 438–444, 1999.

[16] R. P. W. Duin, E. Pekalska, and D. de Ridder, "Relational discriminant analysis," *Pattern Recognition Letters*, vol. 20, pp. 1175–1181, 1999.