Multi-task Regularization of Generative Similarity Models

Luca Cazzanti[1], Sergey Feldman[2], Maya R. Gupta[2], and Michael Gabbay[1]

[1] Applied Physics Laboratory and
[2] Dept. Electrical Engineering
University of Washington
Seattle, USA.

Abstract. We investigate a multi-task approach to similarity discriminant analysis, where we propose treating the estimation of the different class-conditional distributions of the pairwise similarities as multiple tasks. We show that regularizing these estimates together using a least-squares regularization weighted by a task-relatedness matrix can reduce the resulting maximum a posteriori classification errors. Results are given for benchmark data sets spanning a range of applications. In addition, we present a new application of similarity-based learning to analyzing the rhetoric of multiple insurgent groups in Iraq. We show how to produce the necessary task relatedness information from standard given training data, as well as how to derive task-relatedness information if given side information about the class relatedness.

Keywords: similarity, generative similarity-based classification, discriminant analysis, multi-task learning, regularization

1 Introduction

Generative classifiers estimate class-conditional distributions from training samples, and then label a new sample as the class most likely to have generated it [21]. In standard metric-space learning problems, the class-conditional distributions' support is over the Euclidean space of feature vectors. For example, a standard metric-space generative classifier is quadratic discriminant analysis (QDA), which models each class as a multivariate Gaussian [16,35]. More flexible generative models include Gaussian mixture models [19,21] and locally Gaussian models [17].

In contrast, generative similarity-based classifiers build class-conditional probabilistic models of the similarities between training samples, and label any new sample as the class most likely to have generated its similarities to the training data. That is, in generative similarity-based classification, the class-conditional distributions' support is over a similarity space. Similarity discriminant analysis (SDA) models the class-conditional distributions of the similarities as exponential functions [7]. The local similarity discriminant classifier (local SDA) models as exponential functions the class-conditional distributions of the similarities of

a test sample to the k-most similar samples from a training set [5]. Successful classification with local SDA, as with any generative similarity-based or feature-based classifier, depends on the ability to estimate numerically-stable model parameters. A standard approach to ensuring low variance parameter estimates is regularization.

This paper proposes a multi-task approach to regularizing the parameters of the class-conditional exponential models in the local SDA classifier. The motivating hypothesis of the multi-task approach is that learning multiple related tasks in parallel can reduce estimation variance when compared to learning the tasks individually. The successful application of the multi-task approach to many different problems empirically supports this hypothesis, as we briefly review in Sect. 4.

In this paper, the individual tasks consist of estimating the mean of the similarities between samples from pairs of classes. The standard single-task local SDA classifier estimates each of these class-conditional mean similarities independently. In the proposed multi-task approach, the mean estimates are regularized toward each other proportionally to their degree of relatedness, which is captured by a task relatedness matrix. The multi-task regularized mean estimates produce more robust local SDA exponential models which result in improved classification.

Our focus in this paper is on multi-task regularization for SDA. However, SDA is only one of many possible methods for similarity-based learning. Besides local SDA, a different approach to generative classification based on pairwise similarities treats the vector of similarities between any sample and the training set as a feature vector and applies standard feature-space generative classifiers to the similarities-as-features. A drawback of this approach is that the model complexity grows linearly or exponentially with the size of the training set [28, 29]. Other related research considers generative models for graphs [20], where a graph is modeled as being drawn from an exponential distribution.

Other similarity-based learning methods are not generative. Nearest-neighbor methods mirror standard metric space classifiers such as k-nearest neighbors (k-NN) and classify objects based on their most similar neighbors from a training set. Discriminative approaches to similarity-based classification also exist, owing to the popularity of kernel methods such as support vector machines (SVMs). One such approach treats the similarities as features, and mirrors the standard SVM trick of forming kernels by way of inner products (or exponential functions) operating on the vectors of similarities [18, 24].

Another approach treats the entire matrix of training pairwise similarities as the kernel. Since similarities are more general than inner products, the given similarity matrix may be indefinite and must be transformed into an admissible positive semi-definite kernel for use with an SVM [8–10, 26, 31, 37, 38]. SVM-KNN is a local SVM classifier that trains the SVM only on a test sample's k-most similar neighbors in similarity space [39]. For indefinite similarities, it was found to be advantageous to use local similarity-based classifiers such as SVM-KNN or kernel ridge interpolation weighted k-NN that approximate the

local similarity matrix with a positive definite matrix, because lower-error matrix approximations are needed for local neighborhoods than for the entire matrix. For a recent, comprehensive review of similarity-based classifiers, see Chen et al. [9].

In Sect. 2 we briefly review the necessary background on local SDA and illustrate how the need for regularization arises. Section 3 introduces the proposed multi-task regularization for local SDA, shows that the regularized mean similarities have a closed-form solutions and discusses possible choices for the task relatedness matrix. Section 4 discusses other approaches to multi-task learning and contrasts them to the proposed approach. Section 5 reports experimental results on a set of benchmark similarity datasets spanning many different types of similarities, and Sect. 6 reports the results for a document classification problem where the documents are transcripts of statements made by Iraqi insurgent groups. Section 7 concludes with some open questions.

2 Background on Local Similarity Discriminant Analysis

Local SDA models the distribution of similarities as discrete exponentials, and takes its name from the discriminant curves that form the class boundaries in similarity space, in analogy to the standard feature-space classifier QDA, which forms discriminants in feature space. Also in analogy with feature-space generative classifiers, local SDA follows from the standard a posteriori Bayes classifier, which assigns a class label to a test sample x according to the rule

$$y = \arg\max_g P(x|Y=g)P(Y=g)\,,$$

where $P(x|Y=g)$ is the probability of x having been generated from class g, and $P(Y=g)$ is the class g prior probability.

For similarity-based classification, assume that the test and training samples belong to an abstract space $\mathcal{B}$, such as the set of available Internet downloads, the set of amino acid sequences, or the set sonar echoes. 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 $r = |\Omega|$. Alternatively, the pairwise similarity for all training and test samples considered could be given. 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 standard local SDA classifier makes the fundamental assumption that all the information about x's class label depends only on a set of local similarity statistics computed from $\mathcal{N}(x)$, $\mathcal{T}(x) = \bigcup_{h=1\ldots G} T_h(x)$, where $T_h(x)$ is the local similarity statistic computed from $\mathcal{N}_h(x)$. Given a test sample x, the local SDA

classifier assigns x the label y according to the maximum a posteriori rule

$$y = \arg\max_g P(\mathcal{T}(x)|Y=g)P(Y=g) \tag{1}$$

$$= \arg\max_g \prod_{h=1}^{G} P_h(T_h(x)|Y=g)P(Y=g), \tag{2}$$

where (2) is produced from (1) by assuming that the similarity statistics are independent such that the joint class-conditional probability is the product of the marginals.

Several choices are possible for the local similarity statistics $\mathcal{T}(x)$ [5, 7]. In practice, an effective choice are the sets of similarities between x and all its k most-similar neighbors from each class [32], so that $\mathcal{T}_h(x) = \{s(x,z)|z \in \mathcal{N}_h(x)\}$. With this choice, each class-conditional marginal pmf is modeled as the average of exponential functions of the similarities:

$$P_h(T_h(x)|Y=g) \stackrel{\triangle}{=} \frac{1}{k_h} \sum_{z \in \mathcal{N}_h(x)} \hat{P}(s(x,z)|Y=g)$$

$$= \frac{1}{k_h} \sum_{z \in \mathcal{N}_h(x)} \gamma_{gh} e^{\lambda_{gh} s(x,z)}, \tag{3}$$

where $k_h = |\mathcal{N}_h(x)|$.

Each of the parameters $\{\lambda_{gh}\}$ is determined by numerical minimization under the following method-of-moments constraint that the expected value of the similarity be equal to the average similarity computed from the neighborhood training samples:

$$E_{P_h(\mathcal{T}_h(x)|Y=g)}[s(X,z)] = \frac{\sum_{z_a \in \mathcal{N}_g(x)} \sum_{z_b \in \mathcal{N}_h(x)} s(z_a, z_b)}{k_g k_h}. \tag{4}$$

Each of the G^2 mean-constraints (4) is solved by one unique λ_{gh}, but there may be numerical difficulties. For example, when the neighborhood is small, when the discrete similarity domain consists of few distinct values, or when all of x's neighbors are equally maximally (or minimally) similar to each other, the local mean constraint could take on an extremal value:

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

There is no solution to (5) with a finite λ_{gh} – the solution to (5) is a Kronecker delta pmf, $\delta(s(x,z) - c)$. In practice, degenerate pmfs can also arise when a feasible solution exists, but give rise to an exponential function so steep that it effectively acts like a Kronecker delta, which incorrectly concentrates all probability mass on an extremal similarity value c, causing classification errors.

The local SDA formulation (3) mitigates, but does not eliminate, the deleterious effects of degenerate pmfs by modeling the class-conditional marginals as

averages of exponentials, which smooth – or regularize – the effects of the components [32]. Other strategies considered in previous work included regularizing the exponential pmfs by convex combinations with averages of local exponential pmfs, and regularizing the model parameters (the exponents or the means) by convex combinations with baseline parameter values [6]. Yet another strategy considered a Bayesian estimation approach whereby the requirement that the pmfs be exponential was relaxed and the similarities were assumed multinomially distributed with Dirichlet priors on the parameters [32].

All previous strategies regularized each class-conditional pmf in isolation. In the following we present the main contribution of this paper: a multi-task strategy for regularizing the pmfs that exploits the relatedness between the classes.

3 Multi-task Regularization of Mean Similarity Estimates

Given a test sample x, we define as one task the problem of estimating the $(g\,,h)$ mean class-conditional pairwise similarity that appears on the right-hand-side of (4) to solve for the local exponential model. As discussed in the previous section, simply taking the empirical average can lead to numerical problems and non-finite estimates for λ_{gh}. Instead, we propose estimating all G^2 mean class-conditional pairwise similarities jointly as a multi-task problem. Then we use the $(g\,,h)$ multi-task estimate as the right-hand side of (4) to solve for a more stable exponential class-conditional model.

Denote the set of G^2 average similarities by $\{v_{gh}\}$, where v_{gh} is the average similarity between x's neighbors that belong to class g and x's neighbors that belong to class h. That is, $\{v_{gh}\}$ are the average similarities the on right side of (4). We find regularized estimates of the mean similarities

$$\{v^*_{gh}\}_{g,h=1}^{G} = \underset{\{\hat{v}_{gh}\}_{g,h=1}^{G}}{\arg\min} \sum_{g,h=1}^{G} \sum_{z_a \in \mathcal{N}_g(x)} \sum_{z_b \in \mathcal{N}_h(x)} (s(z_a, z_b) - \hat{v}_{gh})^2) + \eta \sum_{j,k=1}^{G} \sum_{l,m=1}^{G} A(v_{jk}, v_{lm})(\hat{v}_{jk} - \hat{v}_{lm})^2. \quad (6)$$

Substituting the solutions into the mean constraint equations (4) yields the regularized-mean constraints

$$E_{P_h(\mathcal{T}_h(x)|Y=g)}[s(X,z)] = v^*_{gh}\,, \quad (7)$$

whose numerical solutions produce the corresponding local SDA model parameters $\{\lambda^*_{gh}\}$.

The first term of (6) minimizes the empirical loss. If one solves (6) with no regularization ($\eta = 0$), the solutions are simply the empirical average similarities $\{v_{gh}\}$. The second term of (6) regularizes the average similarities proportionally to their degree of relatedness, which is captured by the $G^2 \times G^2$ matrix A. Each element $A(v_{jk}, v_{lm})$ quantifies the relatedness of the tasks. We base the

task relatedness on the empirical average similarities v_{jk} and v_{lm}. We detail our choice for the relatedness A in Sect. 3.2.

The regularizing action of the second term of (6) shrinks the mean similarities toward each other, but weights the shrinkage by their relatedness. The effect in the degenerate case (5) is that the average similarity moves away from the extremal value c and shrinks toward the average similarity estimates for the other pmfs proportionally to their relatedness. Thus, the corresponding exponential class-conditional pmf estimate becomes feasible, that is the average similarity has been regularized.

Note that the regularization operates across classes: The average similarity of samples from class g to samples of class h, v_{gh}, is regularized toward the average similarity of the samples from class l to class m, v_{lm}. This is in contrast with other multi-task learning approaches, which associate a task with a sample; instead, the proposed approach associates each task to an exponential class-conditional marginal pmf, which is uniquely determined by the average local similarity parameter. Thus, matrix A captures the degree of relatedness between two exponential pmfs.

3.1 Closed-form Solution

The minimization problem in (6) is convex and, if A is invertible, has the closed-form solution

$$v^* = (I - \tilde{A})^{-1}\tilde{v}\,, \tag{8}$$

where I is the diagonal unit matrix. The vector $\tilde{v} \in \mathbb{R}^{G^2}$ and the matrix $\tilde{A} \in \mathbb{R}^{G^2 \times G^2}$ have components:

$$\tilde{v}_{gh} = \frac{\sum_{z_a \in \mathcal{N}_g(x)} \sum_{z_b \in \mathcal{N}_h(x)} s(z_a, z_b)}{k_g k_h + \eta \sum_{l,m \neq g,h} A(v_{gh}, v_{lm})} \text{ and}$$

$$\tilde{A}(v_{gh}, v_{lm}) = \begin{cases} \frac{\eta A(v_{gh}, v_{lm})}{k_g k_h + \eta \sum_{g,h \neq l,m} A(v_{gh}, v_{lm})} & \text{for} \{g,h\} \neq \{j,k\} \\ 0 & \text{for} \{g,h\} = \{j,k\} \end{cases}$$

These expressions result from setting to zero the partial derivatives of (6) with respect to $\hat{v}_{gh}$, assuming that the task relatedness A is symmetric, and simplifying.

3.2 Choice of Task Relatedness A

Ideally, the task relatedness matrix A conveys information about the strength of the connection between the tasks, but any symmetric invertible matrix can be used as the task relatedness matrix A. For the benchmark classification experiments in Sect. 5, we define A using a Gaussian kernel operating on the differences of the average similarities,

$$A(v_{jk}, v_{lm}) = e^{-(v_{jk} - v_{lm})^2/\sigma}\,. \tag{9}$$

The choice of the Gaussian kernel for A in (6) has an intuitive interpretation. When the average similarities v_{jk} and v_{lm} are close to each other (in the squared difference sense), the Gaussian kernel weights their contribution to the regularization more heavily. When the average similarities are far apart, their reciprocal regularizing influence is greatly diminished by the exponential decay of the Gaussian. The effect is to emphasize the reciprocal influence of closely related average similarities and to discount unrelated mean values, thus preventing unrelated tasks from introducing undue bias in the regularized estimates.

More generally, the task affinity may be mathematically-poorly-defined domain knowledge about how the classes in a particular problem relate to each other. For example, in the insurgent rhetoric classification problem of Sect. 6, we use side information to produce A based on a measure of relatedness between groups that is proportional to number of communiqués jointly released by insurgent groups. The proposed approach can flexibly incorporate such a priori side information about the tasks in the form of matrix A.

4 Related Work in Multi-task Learning

Many new multi-task learning (MTL) methods have been proposed and shown to be useful for a variety of application domains [1, 2, 4, 13, 14, 23, 25, 34, 40]. Such methods comprise both discriminative and generative approaches that either learn the relatedness between tasks or, like this work, assume that a task-relatedness matrix is given.

Recently, multi-task learning research has focused on the problem of simultaneously learning multiple parametric models like multiple linear regression tasks and multiple Gaussian process regression [2, 4, 14]. Some of these multi-task methods jointly learn shared statistical structures, such as covariance, in a Bayesian framework [4]. Zhang and Yeung [40] assumed there exists a (hidden) covariance matrix for the task relatedness, and proposed a convex optimization approach to estimate the matrix and the task parameters in an alternating way. They develop their technique from a probabilistic model of the data and extend it to kernels by mapping the data to a reproducing kernel Hilbert space.

For SVMs, multi-task kernels have been defined [27].Evgeniou et al. [13] proposed a MTL framework for kernels that casts the MTL problem as a single-task learning problem by constructing a special single kernel composed of the kernels from each task. The tasks are learned and regularized simultaneously.

Sheldon [33] builds on the work of Evgeniou et al. [13] and proposes a graphical multi-task learning framework where the tasks are nodes in a graph and the task relatedness information is captured by a kernel defined as the pseudoinverse of the weighted graph Laplacian. This task kernel penalizes distant tasks and shrinks more related tasks toward each other, but in practice must itself be regularized to avoid overfitting. The concept of a task network is also taken up by Kato et al. [23], who combine it with local constraints on the relatedness of pairs of tasks in a conic programming formulation to simultaneously solve for the tasks using kernel machines.

A recent approach integrates semisupervised learning with multi-task learning [25]. In that work both unlabeled and labeled data contribute to the simultaneous estimation of multiple tasks, and their contribution is weighted by their pairwise similarity,which is taken to be a radial basis kernel defined on the difference between feature vectors. We will not discuss in detail here related work in domain adaptation methods and transfer learning [11], which we differentiate as methods that compute some estimates for some tasks, and then regularize estimates for new tasks to the previous tasks' estimates.

The major difference between the existing and the proposed MTL approaches is that the existing approaches do not target similarity-based classifiers. The natural support for existing MTL methods is the Euclidean feature space, and adapting them to similarity-based learning remains an open question beyond the scope of this paper. In contrast, the proposed multi-task regularization naturally operates in similarity space and is ideally suited for generative similarity-based classifiers such as local SDA. Furthermore, as we discuss in Sect. 7, the proposed multi-task regularization approach can be extended to standard Euclidean-space classification and regression tasks.

5 Benchmark Classification Results

We compare the classification performance of the the multi-task regularized local SDA classifiers to the standard single-task local SDA classifier, where the chosen task affinity is the Gaussian kernel operating on the average similarities (9). For comparison, we also report classification results for the k-NN classifier in similarity space and for the SVM-KNN classifier, where the chosen SVM kernel is the inner product of vectors of similarities-as-features. We report classification results for six different benchmark similarity datasets from a variety of applications[1]. More classifier comparisons and details about these datasets can be found in Chen et al. [9].

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 dataset is especially interesting because the similarity function strongly violates the triangle inequality and the minimality property of metrics (a sample should be maximally similar to itself), because customers often buy a different book if they first view a poorly-reviewed book.

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 similarity from 2 to 10 [30]. Perceptual similarities are often non-metric, in that they do not satisfy the triangle inequality.

[1] Datasets and software available at http://staff.washington.edu/lucagc

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 [12]. The self-similarity is set to 1. Like the Amazon dataset, this is a sparse dataset and most of the similarities equal to zero.

The Protein problem is to classify 213 proteins into one of four protein classes based on a sequence-alignment similarity [22].

The Voting problem is to classify 435 representatives into two political parties based on their votes [3]. The categorical feature vector of yes/no/abstain votes was converted into pairwise similarities using the value difference metric, which is a dissimilarity designed to be useful for classification [36]. The voting similarity is a pseudo-metric.

The Face Recognition problem consists of 945 sample faces of 139 people from the NIST Face Recognition Grand Challenge data set. There are 139 classes, one for each person. Similarities for pairs of the original three-dimensional face data were computed as the cosine similarity between integral invariant signatures based on surface curves of the face [15].

The six datasets are divided in 20 disjoint 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. The choice of neighborhood sizes was $\{2, 4, 8, 16, 32, 64, \min(n, 128)\}$. The regularizing parameter η and the kernel bandwidth σ were cross-validated independently of each other among the choices $\{10^{-3}, 10^{-2}, 0.1, 1, 10\}$.

Table 1 shows the mean error rates. Across five datasets multi-task local SDA outperforms single-task local SDA (one dataset is a tie) and for all six datasets it performs better than similarity k-NN. For Sonar and Voting, multi-task local SDA brings the performance closer to SVM-KNN.

Table 1. Percent test error averaged over 20 random test/train splits for the benchmark similarity datasets. Best results are in bold.

	Amazon	Sonar	Patrol	Protein	Voting	FaceRec
	2 classes	2 classes	8 classes	4 classes	2 classes	139 classes
Multi-task Local SDA	**8.95**	14.50	**11.56**	**9.77**	5.52	**3.44**
Local SDA	11.32	15.25	**11.56**	10.00	6.15	4.23
Similarity k-NN	12.11	15.75	19.48	30.00	5.69	4.29[2]
SVM-KNN (sims-as-features)	13.68	**13.00**	14.58	29.65	**5.40**	4.23[2]

[2] Results for k-NN and SVM-KNN were reported previously. The same train/test splits were used, but the cross-validation parameters were slightly different. See Chen et al. for details [9].

6 Iraqi Insurgent Rhetoric Analysis

We address the problem of classifying the rhetoric of insurgent groups in Iraq. The data consist of 1924 documents – translated jihadist websites or interviews with insurgent officials – provided by the United States government's Open Source Center. We consider the problem of classifying each document as having been released by one of eight insurgent groups operating in Iraq from 2003 to 2009.

Each document is represented by a 173-dimensional vector whose elements contain the frequency of occurrence of 173 keywords in the document. The dictionary of keywords was defined by one of the authors, who is an expert on insurgent rhetoric analysis. The chosen document similarity was the symmetrized relative entropy (symmetrized Kullback-Leibler divergence) of the normalized keyword frequency vectors.

For this problem, we compared two definitions of the task relatedness. One, we defined the task relatedness as proposed in Sect. 3. In addition, we derived a task relatedness from side information about the number of communiqués jointly released by two groups, shown in Table 2, where the j-th row and the k-th column denote the number of communiqués jointly released by the j-th and k-th insurgent groups. This side information was derived from a smaller, separate dataset. A higher number of joint statements indicates more cooperation among the leaders of the two groups and, typically, greater ideological affinity as well. Note that some groups work in isolation, while others selectively choose their collaborators.

Table 2. Number of Communiqués Jointly Released By Any Two Groups

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
Group 1	0	0	0	0	7	8	6	2
Group 2	0	0	0	0	0	0	0	0
Group 3	0	0	0	0	0	0	0	0
Group 4	0	0	0	0	0	0	0	1
Group 5	7	0	0	0	0	6	5	1
Group 6	8	0	0	0	6	0	5	1
Group 7	6	0	0	0	5	5	0	1
Group 8	2	0	0	1	1	1	1	0

We conjecture that an appropriate multi-task regularization is to shrink the average document similarity estimates of more strongly connected groups toward each other. Recall that for local SDA there are G^2 mean-similarity constraints, where G is the number of classes. Each constraint is associated with its corresponding task of estimating the class-conditional marginal exponential pmf of

the similarity between documents from group j and documents from group k, and consequently the task relatedness matrix A has dimensions $G^2 \times G^2$. In this problem there are $G = 8$ insurgent groups. Let the 8×8 matrix given in Table 2 be denoted by Q. We form the 64×64 task relatedness A from the joint communiqués Q as

$$A(v_{jk}, v_{lm}) = e^{-(Q_{jk} - Q_{lm})2/\sigma} . \tag{10}$$

This choice of task relatedness implies that the mean document similarities v_{jk} and v_{lm} should be strongly related if the number of joint communiqués released by groups (j, k) is the same as the number released by groups (l, m), and should be weakly related if the numbers differ greatly. Thus *the task relatedness measures the similarity between pairs of insurgent groups*. Furthermore, choosing a Gaussian kernel operating on all possible differences of the entries in matrix Q ensures that A is invertible.

Table 3 shows the leave-one-out cross-validation error rates for single-task, multi-task local SDA, and similarity k-NN. The neighborhood size and the parameters η and σ were cross validated from parameter choices identical to the benchmark datasets. In addition to the task relatedness derived from the side information Q, we tested the Gaussian kernel operating directly on the class-conditional document similarities in (9) without using any side information. For both choices of task relatedness, the performance of multi-task local SDA provides a small gain over the standard local SDA and similarity k-NN.

Table 3. Percent leave-one-out cross-validation classification error for the insurgent rhetoric document classification problem. Best result is in bold

Multi-task Local SDA (w/ joint statements task relatedness)	**52.34**
Multi-task Local SDA (w/ Gaussian kernel task relatedness)	52.75
Local SDA	54.52
Similarity k-NN	53.53
Guessing Using Class Priors	77.91

The communiqués-derived and the mean document similarity-derived task relatedness definitions represent two approaches to capturing the relationships between the insurgent groups. The former approach incorporates mathematically poorly-defined side information about the problem available from a separate data set, while the latter is purely data-driven from the document similarity data. The multi-task local SDA can flexibly accommodate both types of task knowledge. It is interesting that in this experiment both approaches lead to almost identical classification improvement over single-task local SDA.

Finally, many other definitions of document similarity are possible. While choosing the best similarity is an important practical problem, it is beyond the scope of this paper. In any case, the SDA classification framework, single- or multi-task, is independent of the chosen document similarity function, thus can accommodate any future choice of document similarity.

7 Discussion and Open Questions

In this paper, we have proposed treating the estimation of different class-conditional distributions in a generative model as multiple tasks, and shown that regularizing these estimates together with a simple least-squares similarity-based regularization can reduce classification errors.

It can be argued that regularizing the class-conditional distributions toward each other according to their relatedness implies that the class-conditional local SDA models are in fact correlated, which appears inconsistent with the assumptions that the class-conditional marginals in the SDA classifier (2) are independent. It might be possible to model the correlations directly in the SDA model without resorting to multi-task regularization, but this strategy must contend with the concomitant problem of having to estimate the task correlations in addition to the task-specific parameters, and makes the SDA classifier more complex. In constrast, the proposed multi-task regularization does not impose a particular structure on the task relatedness (i. e. correlation), which can be provided as domain-specific knowledge or computed directly – not estimated – from the task-specific parameters. We argue that this approach is more flexible, because it does not require modifying the original classifer, and more general, because it accommodates any problem-relevant task relatedness.

In the SDA model, the class-conditional pmf $P_j(T_j(x)|k)$ models the similarity of samples from class k to the samples of class j. Thus to tie the $P_j(T_j(x)|k)$ task to the $P_m(T_m(x)|l)$ task, we need the relatedness between the pair of classes $(j\,,k)$ to the pair of classes $(l\,,m)$. A simpler model would be to tie together tasks based only on one of the involved classes: Tie together the $P_j(T_j(x)|k)$ and $P_m(T_m(x)|l)$ tasks based only on the relatedness between the k-th and m-th classes or only on the relatedness between the j-th and l-th classes. Then the task-relatedness would simply be the class-relatedness.

Side information about class relatedness could be used, like the group-relatedness given in the group rhetoric analysis in Sect. 6. In the absence of side information, class-relatedness could be produced by first running a single-task classifier (like local SDA) and using the resulting class-confusion matrix as the task-relatedness matrix for the multi-task classifier. However, an advantage to the approach we took here of tying pairs of classes together is that we use the relatedness of both the (j,k) pair and the (l,m) pair, and by using a Gaussian RBF kernel to form A, an invertible A is always produced, ensuring a closed-form solution.

A more general nonparametric multi-task learning formulation would be

$$\{y_t^*\}_{t=1}^U = \underset{\{\hat{y}_t\}_{t=1}^U}{\arg\min} \sum_{t=1}^{U} \sum_{i=1}^{N_t} L(y_{ti}, \hat{y}_t) + \gamma J\left(\{\hat{y}_t\}_{t=1}^T\right), \tag{11}$$

where L is a loss function, J is a regularization function, U is the number of tasks, and N_t is the number of data points from task t. However, an advantage of the squared error formulation given in (6) is that is has a closed-form solution, as given in Sect. 3.1.

A number of theoretical questions can be asked about the proposed multi-task framework. Many MTL methods have a Bayesian interpretation, in that the task-specific random variables can be modeled as drawn from some shared prior, such that joint shrinkage towards the mean of that prior is optimal. In our cases, however, the shrinkages are mutual, and we hypothesize that an empirical Bayesian perspective would be needed. Ideally, the assumed multi-task similarities would perfectly represent the underlying statistical relatedness of the tasks. For what types of statistical relatedness is the proposed multi-task learning optimal, and what would the corresponding optimal task relatedness look like? Further, to what extent can one estimate an optimal task relatedness matrix of interest from the statistics of the tasks, with or without side information?

References

1. Agarwal, A., Daumé III, H., Gerber, S.: Learning multiple tasks using manifold regularization. In: Lafferty, J., Williams, C.K.I., Shawe-Taylor, J., Zemel, R., Culotta, A. (eds.) Advances in Neural Information Processing Systems 23, pp. 46–54 (2010)
2. Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. Machine Learning 73(3), 243–272 (2008)
3. Asuncion, A., Newman, D.J.: UCI machine learning repository (2007), `http://www.ics.uci.edu/$\sim$mlearn/{MLR}epository.html`
4. Bonilla, E.V., Chai, K.M.A., Williams, C.K.I.: Multi-task Gaussian process prediction. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (eds.) Advances in Neural Information Processing Systems. MIT Press, Cambridge, MA (2008)
5. Cazzanti, L., Gupta, M.R.: Local similarity discriminant analysis. In: Proc. Intl. Conf. Machine Learning (2007)
6. Cazzanti, L., Gupta, M.R.: Regularizing the local similarity discriminant analysis classifier. In: Proc. 8th Intl. Conf. Machine Learning and Applications (December 2009)
7. Cazzanti, L., Gupta, M.R., Koppal, A.J.: Generative models for similarity-based classification. Pattern Recognition 41(7), 2289–2297 (July 2008)
8. Chen, J., Ye, J.: Training svm with indefinite kernels. Proc. of the Intl. Conf. on Machine Learning (2008)
9. Chen, Y., Garcia, E.K., Gupta, M.R., Rahimi, A., Cazzanti, L.: Similarity-based classification: Concepts and algorithms. Journal of Machine Learning Research 10, 747–776 (March 2009)
10. Chen, Y., Gupta, M.R.: Learning kernels from indefinite similarities. Proc. of the Intl. Conf. on Machine Learning (2009)
11. Daumé III, H., Marcu, D.: Domain adaptation for statistical classifiers. Journal of Artificial Intelligence Research 26, 101–126 (2006)
12. Driskell, J.E., McDonald, T.: Identification of incomplete networks. Florida Maxima Corporation Technical Report (08–01) (2008)
13. Evgeniou, T., Micchelli, C., Pontil, M.: Learning multiple tasks with kernel methods. Journal of Machine Learning Research (6) (April 2005)
14. Evgeniou, T., Pontil, M.: Regularized multi–task learning. In: KDD '04. pp. 109–117. ACM, New York, NY, USA (2004)

15. Feng, S., Krim, H., Kogan, I.A.: 3D face recognition using Euclidean integral invariants signature. In: Statistical Signal Processing, 2007. SSP '07. IEEE/SP 14th Workshop on (2007)
16. Friedman, J.H.: Regularized discriminant analysis. Journal American Statistical Association 84(405), 165–175 (1989)
17. Garcia, E.K., Feldman, S., Gupta, M.R., Srivastava, S.: Completely lazy learning. IEEE Trans. Knowledge and Data Engineering 22(9), 1274–1285 (2010)
18. Graepel, T., Herbrich, R., Bollmann-Sdorra, P., Obermayer, K.: Classification on pairwise proximity data. In: Advances in Neural Information Processing Systems. vol. 11, pp. 438–444 (1998)
19. Gupta, M.R., Chen, Y.: Theory and use of the em method. Foundations and Trends in Signal Processing 4(3), 223–296 (2010)
20. Handcock, M., Hunter, D.R., Goodreau, S.: Goodness of fit of social network models. In: Journal American Statistical Association. vol. 103, pp. 248–258 (2008)
21. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer-Verlag, New York (2001)
22. Hofmann, T., Buhmann, J.: Pairwise data clustering by deterministic annealing. IEEE Trans. on Pattern Analysis and Machine Intelligence 19(1) (January 1997)
23. Kato, T., Kashima, H., Sugiyama, M., Asai, K.: Multi-task learning via conic programming. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (eds.) Advances in Neural Information Processing Systems 20, pp. 737–744. MIT Press, Cambridge, MA (2008)
24. Liao, L., Noble, W.S.: Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. Journal of Computational Biology 10(6), 857–868 (2003)
25. Liu, Q., Liao, X., Li, H., Stack, J.R., L.Carin: Semisupervised multitask learning. IEEE Trans. Pattern Analysis and Machine Intelligence (6) (June 2009)
26. Luss, R., d'Aspremont, A.: Support vector machine classification with indefinite kernels. Mathematical Programming Computation 1(2), 97–118 (2009)
27. Micchelli, C.A., Pontil, M.: Kernels for multi–task learning. In: Advances in Neural Information Processing Systems (2004)
28. Pekalska, E., Duin, R.P.W.: Dissimilarity representations allow for building good classifiers. Pattern Recognition Letters 23(8), 943–956 (June 2002)
29. Pekalska, E., Paclíc, P., Duin, R.P.W.: A generalized kernel approach to dissimilarity-based classification. Journal of Machine Learning Research pp. 175–211 (2001)
30. Philips, S., Pitton, J., Atlas, L.: Perceptual feature identification for active sonar echoes. In: Proc. of the 2006 IEEE OCEANS Conf. (2006)
31. Roth, V., Laub, J., Kawanabe, M., Buhmann, J.M.: Optimal cluster preserving embedding of nonmetric proximity data. IEEE Trans. Pattern Anal. and Machine Intel. 25(12), 1540–1551 (Dec 2003)
32. Sadowski, P., Cazzanti, L., Gupta, M.R.: Bayesian and pairwise local similarity discriminant analysis. In: Proc. Intl. Workshop on Cognitive Information Processing (CIP). Isola d'Elba, Italy (June 2010)
33. Sheldon, D.: Graphical multi-task learning (2008), `http://web.engr.oregonstate.edu/~sheldon`, neural Information Processing Systems Workshops
34. Sheldon, D.: Graphical multi-task learning (2010), `http://web.engr.oregonstate.edu/~sheldon`, unpublished manuscript
35. Srivastava, S., Gupta, M.R., Frigyik, B.: Bayesian quadratic discriminant analysis. Journal of Machine Learning Research 8, 1277–1305 (2007)

36. Stanfill, C., Waltz, D.: Toward memory-based reasoning. Communications of the ACM 29(12), 1213–1228 (December 1986)
37. Wu, G., Chang, E.Y., Zhang, Z.: An analysis of transformation on non-positive semidefinite similarity matrix for kernel machines. Tech. rep., University of California, Santa Barbara (March 2005)
38. Ying, Y., Campbell, C., Girolami, M.: Analysis of svm with indefinite kernels. In: Advances in Neural Information Processing Systems. MIT Press, Cambridge, MA (2009)
39. Zhang, H., Berg, A.C., Maire, M., Malik, J.: 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)
40. Zhang, Y., Yeung, D.Y.: A convex formulation for learning task relationships. In: Grünwald, P., Spirtes, P. (eds.) Proc. of the 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010) (2010)