FINE: Fisher Information Nonparametric Embedding

Kevin M. Carter, *Student Member*, *IEEE*,
Raviv Raich, *Member*, *IEEE*, William G. Finn,
and Alfred O. Hero III, *Fellow*, *IEEE*

Abstract—We consider the problems of clustering, classification, and visualization of high-dimensional data when no straightforward euclidean representation exists. In this paper, we propose using the properties of information geometry and statistical manifolds in order to define similarities between data sets using the Fisher information distance. We will show that this metric can be approximated using entirely nonparametric methods, as the parameterization and geometry of the manifold is generally unknown. Furthermore, by using multidimensional scaling methods, we are able to reconstruct the statistical manifold in a low-dimensional euclidean space; enabling effective learning on the data. As a whole, we refer to our framework as Fisher Information Nonparametric Embedding (FINE) and illustrate its uses on practical problems, including a biomedical application and document classification.

Index Terms—Information geometry, statistical manifold, dimensionality reduction, multidimensional scaling.

1 Introduction

Due to the ever expanding capabilities of data retrieval, data are often represented in some high-dimensional fashion, leading to difficulties in learning tasks due to the *curse of dimensionality*. While the problem of learning in an euclidean space has been thoroughly researched in manifold learning, there are many problems in which the data cannot be appropriately represented as a (Riemannian) submanifold of the euclidean space, and the model parameters are unspecified and must be learned through the data. In contrast to the ad hoc "solution" of processing the data as real-valued feature vectors in the euclidean space, we consider the case that generative models for the data can be represented as points on a *statistical* manifold—a manifold of probability density functions (PDFs). Applications of statistical manifolds have been presented in the cases of document classification [1], face recognition [2], texture segmentation [3], image analysis [4], and clustering [5], all proposing alternatives to using euclidean geometry for data modeling.

When the parameterization of the statistical manifold is available, one can project the data onto the manifold to obtain a corresponding statistical model and the exact geodesic distance can be computed to measure the distance between PDFs. In many problems of practical interest, however, the manifold geometry is unavailable and the calculation of geodesics must be done in a model-free, nonparametric fashion. In this paper, we present a framework—deemed *Fisher Information Nonparametric Embedding* (FINE)—to deal with such problems. FINE includes characterization of data sets in terms of a nonparametric statistical model, a geodesic approximation of the Fisher information distance as a metric for evaluating similarities between data sets, and a dimension reduction procedure to obtain a low-dimensional euclidean embedding of the original high-dimensional data for various learning tasks. Unlike previous presentations of statistical manifolds, our method is entirely nonparametric and contains no model assumptions, yielding a low-dimensional embedding based entirely on the information geometry of the samples.

This paper is organized as follows: Section 2 gives the formulation for the problem we wish to solve, while Section 3 develops and outlines the FINE algorithm. We illustrate the results of using FINE on real data sets in Section 4. Finally, we draw conclusions and discuss the possibilities for future work in Section 5.

2 Problem Formulation

Recent methods of manifold learning and dimensionality reduction [6], [7], [8] focus on finding a low-dimensional representation of the data which are restricted to lie on a (Riemannian) submanifold of an euclidean space. These methods are designed to optimally reconstruct such a euclidean manifold given only a set of sample points which lie on said manifold. While each method implements this optimization differently (i.e., locally, globally, etc.), all are designed to preserve some measure of the L_2 distance norm between sample points in a given data set.

Rather than focusing on reducing the dimension of a single data set to reconstruct its euclidean manifold, we extend the problem to statistical manifolds, or PDFs. In this case, the space in which the data lie is of no particular interest. Instead, the generative model of the data is restricted to lie on a statistical manifold, and we wish to reconstruct the manifold given a collection of sample data *sets*.

Multidimensional scaling (MDS) [9] and its derivations utilize pairwise euclidean distances to recreate manifolds and embed points into a low-dimensional space. However, it has been well documented that these methods use euclidean distance as a measure of dissimilarity between elements, and other measures of dissimilarity may be substituted. Isomap [7], for example, approximates the geodesic distance between data samples. Laplacian eigenmaps [6] simply use euclidean distance as a means to calculate a weight function. Hence, if an appropriate distance between PDFs is utilized, these well-respected algorithms could be used for an entirely new class of problems.

Let $\mathcal{P} = \{p_1, \ldots, p_N\}$ be a collection of PDFs lying on some statistical manifold $\mathcal{M}$. Our goal is to reconstruct $\mathcal{M}$ using only the information available in $\mathcal{P}$. Hence, we would like to find a distance measure between PDFs to calculate the pairwise dissimilarities. This enables the usage of MDS methods to reconstruct a low-dimensional embedding of the statistical manifold in euclidean space. This allows for effective learning on the family of distributions lying on the manifold.

3 Methods

3.1 Fisher Information Distance

For a parametric family of probability distributions on a statistical manifold, it is possible to define a Riemannian metric using the Fisher information matrix $[\mathcal{I}(\theta)]$, which measures the amount of information a random variable contains in reference to an unknown parameter θ. The Fisher information distance between two distributions $p(x;\theta_1)$ and $p(x;\theta_2)$ is

- *K.M. Carter and A.O. Hero, III, are with the Department of Electrical Engineering and Computer Science, University of Michigan, 1301 Beal Ave., Ann Arbor, MI 48109.*
  E-mail: kmcarter@umich.edu, hero@eecs.umich.edu.
- *R. Raich is with the School of EECS, Oregon State University, Kelley Engineering Center 3009, Corvallis, OR 97331.*
  E-mail: raich@eecs.oregonstate.edu.
- *W.G. Finn is with the Department of Pathology, University of Michigan, 1301 Catherine St., Ann Arbor, MI 48109. E-mail: wgfinn@umich.edu.*

Manuscript received 24 Sept. 2008; revised 29 Jan. 2009; accepted 9 Mar. 2009; published online 19 Mar. 2009.
Recommended for acceptance by D.D. Lee.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2008-09-0644.
Digital Object Identifier no. 10.1109/TPAMI.2009.67.

$$D_F(\theta_1,\theta_2) = \min_{\substack{\theta(\cdot):\\ \theta(0)=\theta_1\\ \theta(1)=\theta_2}} \int_0^1 \sqrt{\left(\frac{d\theta}{dt}\right)^T [\mathcal{I}(\theta)] \left(\frac{d\theta}{dt}\right)}\, dt, \tag{1}$$

where $\theta = \theta(t)$ is the parameter path along the manifold [10], [11]. Note that the coordinate system of a statistical manifold is the same as the parameterization of the PDFs (i.e., θ). Essentially, (1) amounts to finding the length of the shortest path—the geodesic—on $\mathcal{M}$ connecting coordinates θ_1 and θ_2.

While the Fisher information distance cannot be exactly computed without a priori knowledge about the geometry (i.e., parameterization) of the manifold, the distance between PDFs p_1 and p_2 may be approximated with a variety of metrics such as the Kullback-Leibler (KL) divergence,

$$KL(p_1 \parallel p_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)}\, dx, \tag{2}$$

the Hellinger distance,

$$D_H(p_1,p_2) = \sqrt{\int \left(\sqrt{p_1(x)} - \sqrt{p_2(x)}\right)^2 dx}, \tag{3}$$

and the cosine distance,

$$D_C(p_1,p_2) = 2 \arccos \int \sqrt{p_1(x) \cdot p_2(x)}\, dx, \tag{4}$$

all of which converge to the Fisher information distance,

$$\begin{aligned} \sqrt{2KL(p_1 \parallel p_2)} &\to D_F(p_1,p_2),\\ 2D_H(p_1,p_2) &\to D_F(p_1,p_2),\\ D_C(p_1,p_2) &\to D_F(p_1,p_2), \end{aligned}$$

as $p_1 \to p_2$ [10]. These measures allow for the approximation of the information distance in the absence of the geometry of the statistical manifold on which the PDFs lie. Note that there exists a monotonic transformation function relating the Hellinger distance to the cosine distance, $\psi : D_H \to D_C$. Additionally, while the KL-divergence is not a symmetric measure, we can add symmetry by defining, $D_{KL}(p_1,p_2) = KL(p_1 \parallel p_2) + KL(p_2 \parallel p_1)$, which maintains the convergence properties. For additional measures of probabilistic distance, some of which approximate the Fisher information distance, and a means of calculating them between data sets, we refer the reader to [12], [13].

It has previously been suggested [3] to use the cosine distance as a strict approximation of the Fisher information distance. This is due to the fact that the cosine distance measures a portion of a great circle on a hypersphere, and in the discrete case, all PDFs can be considered as multinomial distributions which may be projected onto a hypersphere manifold. This usage of the cosine distance is true only in the assumption that the manifold of interest fills the entire space of the hypersphere. In many cases, the PDFs are constrained to form a submanifold of interest, and the geodesic is no longer accurately described as a portion of a great circle on the hypersphere. This is illustrated in Fig. 1 in which we represent a $(d-1)$-dimensional submanifold which occupies a subspace of the d-dimensional hypersphere ($d = 2$ for illustration). The Fisher information distance is equal to the shortest path along the submanifold (curvy line) and is not equal to the portion of a great circle on a hypersphere connecting the two points. Hence, there are situations in which standard approximations of the information distance do not converge to the true distance and it is necessary to approximate the geodesic along the manifold.

Using a graphical model, we may define the path between p_1 and p_2 as a series of connected segments. The geodesic distance may then be approximated as the sum of the length of those segments. Specifically, given the collection of N PDFs $\mathcal{P} = \{p_1, \ldots, p_N\}$ and using an approximation of the Fisher information distance $\hat{D}_F(p_1,p_2)$ as $p_1 \to p_2$, we can now define an approximation function G for all pairs of PDFs:

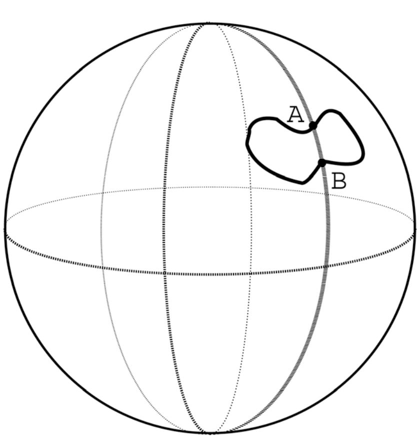


Fig. 1. Given a one-dimensional submanifold (the curvy dark line) of interest lying on a two-dimensional sphere manifold, the Fisher information distance is the shortest path connecting the points A and B along the one-dimensional submanifold, rather than the length of a portion of the great circle connecting the points on the sphere.

$$G(p_1,p_2;\mathcal{P}) = \min_{M,\mathcal{P}} \sum_{i=1}^{M-1} \hat{D}_F\big(p_{(i)}, p_{(i+1)}\big), \quad p_{(i)} \to p_{(i+1)}\ \forall\ i. \tag{5}$$

Intuitively, this estimate calculates the length of the shortest path between points in a connected graph on the well-sampled manifold and, as such, $G(p_1,p_2;\mathcal{P}) \to D_F(p_1,p_2)$ as $N \to \infty$. Empirically, (5) may be solved with Dijkstra's shortest path algorithm. This is similar to the manner in which Isomap [7] approximates manifold distances.

3.2 Dimensionality Reduction

Given a matrix of dissimilarities between entities, many MDS algorithms have been developed to find a low-dimensional embedding of the original data $\psi : \mathcal{M} \to \mathbb{R}^d$. As stated previously, these techniques enable the reconstruction of manifolds from a finite sampling. While historically used to reconstruct euclidean manifolds, we apply the same techniques to reconstruct statistical manifolds by using the Fisher information distance (or approximation thereof) as a pairwise dissimilarity metric. Hence, we are able to find a single low-dimensional coordinate representation of each PDF.

While there are many available MDS methods, in this paper, we utilize classical MDS (cMDS) [9], Laplacian eigenmaps (LEM) [6], and classification constrained dimensionality reduction (CCDR) [14]. It is worth noting that, per our formulation, cMDS operating on the geodesic distance is the same setting as the Isomap algorithm (albeit with a different metric).

3.3 FINE Algorithm

In problems of practical interest, the parameterization of the probability densities is usually unknown. We instead are given a family of data sets $\mathcal{X} = \{X_1, \ldots, X_N\}$, in which we may assume that each data set X_i is a realization of some underlying PDF to which we do not have knowledge of the parameters. Given such problems, nonparametric methods of density estimation such as kernel methods and k-nearest neighbor (k-NN) methods are appropriate to estimate both the PDFs and the approximation of the Fisher information distance. For additional details on both kernel density estimation and

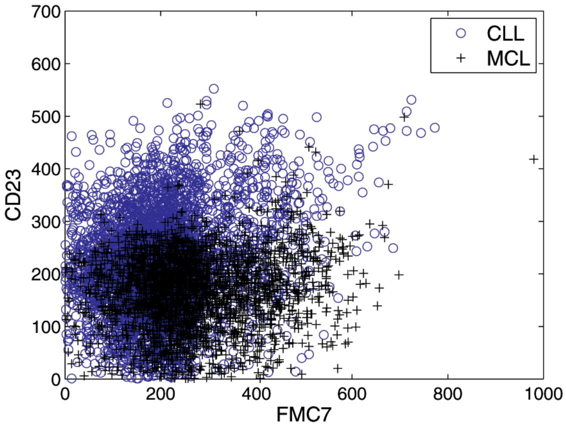


Fig. 2. Two-dimensional plot of patients in disease classes CLL and MCL, in which each point represents a unique blood cell.

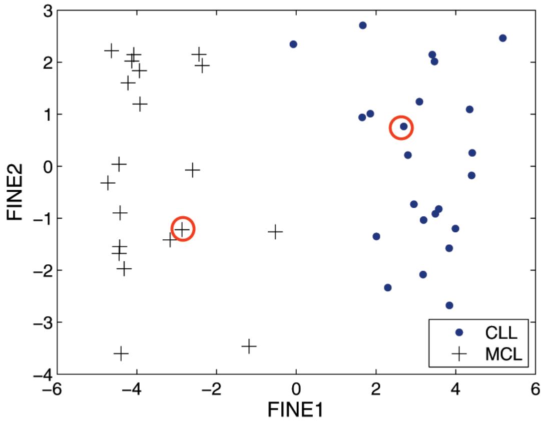


Fig. 3. Two-dimensional embedding of CLL (•) and MCL (+) patients using FINE. The circled points correspond to the CLL and MCL cases illustrated in Fig. 2.

calculation of information divergences, including the specific implementation for our methods, we refer the reader to [13].

FINE is presented in Algorithm 1 and combines the presented methods to find a low-dimensional embedding of a collection of data sets. If we assume that each data set is a realization of an underlying PDF and each of these distributions lie on a statistical manifold with some natural parameterization, then this embedding can be viewed as a reconstruction of the manifold into euclidean space. Note that in line 5, "$\mathrm{mds}(G, d)$" refers to using any multidimensional scaling method to embed the dissimilarity matrix G into a euclidean space with dimension d.

Algorithm 1. Fisher Information Nonparametric Embedding

```
Input: Collection of data sets $\mathcal{X} = \{X_1, \ldots, X_N\}$; the desired
       embedding dimension $d$
1: for $i = 1$ to $N$ do
2:     Calculate $\hat{p}_i(\boldsymbol{x})$, the density estimate of $X_i$
3: end for
4: Calculate $G$, where $G(i, j)$ is the geodesic approximation of the
   Fisher information distance between $p_i$ and $p_j$
5: $Y = \mathrm{mds}(G, d)$
Output: $d$-dimensional embedding of $\mathcal{X}$, into euclidean space
   $\boldsymbol{Y} \in \mathbb{R}^{d \times N}$
```

At this point, it is worth stressing the benefits of this framework. Through information geometry, FINE enables the joint embedding of multiple data sets X_i into a single low-dimensional euclidean space. By viewing each $X_i \in \mathcal{X}$ as a realization of $p_i \in \mathcal{P}$, we reduce the numerous samples in X_i to a single point. The dimensionality of the statistical manifold may be significantly less than that of the euclidean realizations (e.g., a multivariate Gaussian). MDS methods reduce the dimensionality of p_i from the euclidean data dimension to the dimension of the statistical manifold on which it lies. This results in a single low-dimensional representation of each original data set $X_i \in \mathcal{X}$.

4 Applications

We now present practical applications for the FINE framework which are based around visualization and classification. In each application, the densities are unknown, but we assume that they lie on a manifold with some natural parameterization.

4.1 Flow Cytometry

In clinical flow cytometry, pathologists gather readings of fluorescent markers and light scatter off of individual blood cells from a patient sample, leading to a characteristic multidimensional distribution that, depending on the panel of markers selected, may be distinct for a specific disease entity. Clinical pathologists generally interpret results in the form of two-dimensional scatter plots in which the axes each represent one of the many cell characteristics analyzed; the multidimensional nature of flow cytometry is routinely underutilized in practice.

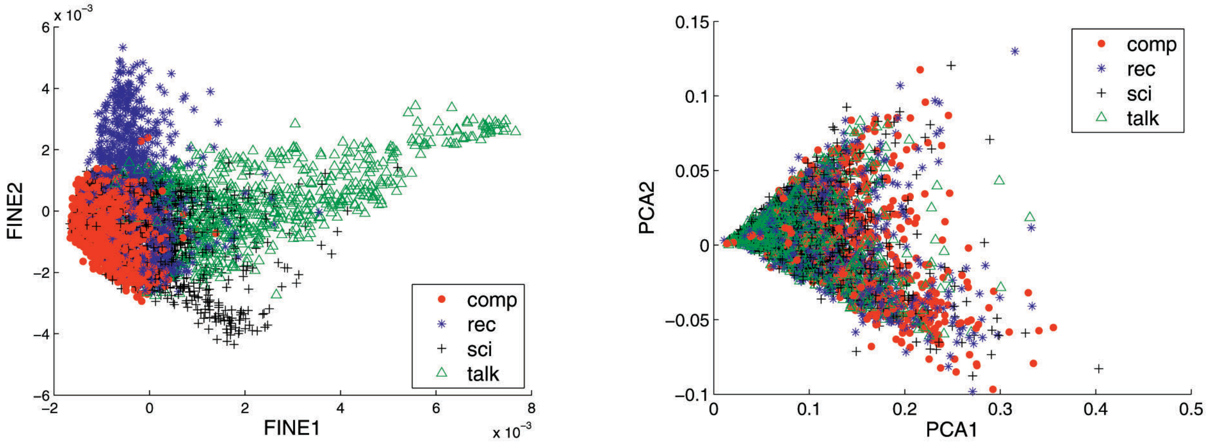
An example of the difficulty in analysis of two-dimensional scatter plots is illustrated in Fig. 2. Two distinct but immunophenotypically similar forms of lymphoid leukemia are shown—mantle cell lymphoma (MCL) and chronic lymphocytic leukemia (CLL). These diseases display similar characteristics with respect to many expressed surface antigens, but are generally distinct in their patterns of expression of two common B lymphocyte antigens CD23 and FMC7. The significant similarity and overlapping nature in the marginal plots illustrates the difficulty in traditional two-dimensional flow cytometry analysis.

While the expression of various markers may be highly variable over different patients, the general characterization of the multivariate PDF underlying each patient sample is much less variable. Hence, each distribution exists on some statistical manifold with a much lower dimensional parameterization, and this application is appropriate for FINE [15], [16]. Specifically, let $\mathcal{X} = \{X_1, \ldots, X_N\}$, where X_i is the data set corresponding to the flow cytometer output of the ith patient. Each patient's blood is analyzed for five parameters: forward and side light scatter, and three fluorescent markers (CD45, CD23, and FMC7). Hence, each data set X_i is five-dimensional with n_i elements corresponding to individual blood cells. Given that $\mathcal{X}$ is comprised both patients with CLL and patients with MCL, we wish to analyze the performance of FINE for the visualization of cytometric data.

The data set[1] consists of 23 patients with CLL and 20 patients with MCL, and the set X_i for each patient is on the order of $n_i \approx 5{,}000$ cells. Densities were approximated with a Gaussian kernel KDE, using the maximal smoothing principal [17] for bandwidth selection. Fig. 3 shows the two-dimensional embedding with FINE, using cMDS and the symmetric KL-divergence set as a local approximation of the Fisher information distance. Each point in the plot represents an entire patient data set. It should be noted that there exists a natural separation between the classes, as the implementation was entirely unsupervised.

An important byproduct of this natural clustering is the ability to visualize the cytometric data in a manner which allows comparisons between patients. The circled points in Fig. 3 correspond to the patients illustrated in Fig. 2, which were difficult to differentiate by using a scatter plot of the most discerning marker combination as

[1] Data and clinical diagnosis for each patient was provided by the Department of Pathology at the University of Michigan.



(a) (b)

Fig. 4. Two-dimensional embeddings of 20 Newsgroups data. The data display some natural clustering in the information-based embedding, while the PCA embedding does not distinguish between classes. (a) FINE. (b) PCA.

deemed by pathologists. In the space defined by FINE, the patients are easily differentiated and lie well within the clusters of each disease type. By using the embedding created with FINE, pathologists are able to visually determine similarities between patients, which gives them a quick and easy means of determining which data sets may need further investigation (e.g., for possible misdiagnosis). For further usages of FINE for flow cytometry analysis, we encourage the reader to view [15], [16].

4.2 Document Classification

Recent work has shown interest in using dimension reduction for the purposes of document classification [18] and visualization [19]. Typically, documents are represented as very high-dimensional PDFs, and learning algorithms suffer from the *curse of dimensionality*. Dimension reduction not only alleviates these concerns, but also reduces the computational complexity of learning algorithms due to the resultant low-dimensional space. As such, the problem of document classification is an interesting application for FINE [20].

Given a collection of documents of known class, we wish to best classify a document of unknown class. A document can be viewed as a realization of some overriding probability distribution on a "bag of words," in which different distributions will generate different documents. In this setting, we defined the PDFs as the *term frequency* representation of each document. Specifically, let x_i be the number of times term i appears in a specific document. The PDF of that document can then be characterized as the multinomial distribution of normalized word counts, with the maximum likelihood estimate provided as

$$\hat{p}(\boldsymbol{x}) = \left(\frac{x_1}{\sum_i x_i}, \ldots, \frac{x_n}{\sum_i x_i} \right), \tag{6}$$

where n is the number of words in dictionary $\boldsymbol{x}$.

For illustration, we will utilize the well-known 20 Newsgroups data set,[2] which contains word counts for 18,774 postings on 20 newsgroups and recommends specific indexes for training and test sets. We choose to restrict our simulation to the four domains with the largest number of subdomains (comp.*, rec.*, sci.*, and talk.*), and classify each posting by its highest level domain. Specifically, we are given $\mathcal{P} = \{p_1, \ldots, p_N\}$ where each p_i corresponds to a single newsgroup posting and is estimated with (6). We note that the data were preprocessed to remove all words that occur in five or less documents.[3]

4.2.1 *Unsupervised FINE*

First, we utilize unsupervised methods to see if a natural separating geometry exists between domains. Using Laplacian eigenmaps on the dissimilarities calculated with the Hellinger distance, we found an embedding $\mathcal{P} \rightarrow \mathbb{R}^2$. Fig. 4a shows the natural geometric clustering between the different document classes, while a principal component analysis (PCA) embedding (Fig. 4b) does not demonstrate the same effect. PCA is often used as a means to lower the dimension of data for learning problems due to its optimality for euclidean data. However, the PCA embedding of the 20 Newsgroups corpus does not exhibit any natural class separation due to the noneuclidean nature of the data.

Extending to document classification, dimensionality reduction is important as the natural dimension (i.e., number of words) for the corpus is 26,214. After embedding $\mathcal{P} \rightarrow \mathbb{R}^d$ in the range $d \in [5, 50]$, we apply a linear kernel support vector machine (SVM) to classify the data in an "all-versus-all" setting (i.e., classify each test sample as one of four different potential classes). The training and test sets were separated according to the recommended indexes, and each set was randomly subsampled for computational purposes, keeping the number of training and test samples constant (400 and 200, respectively). Note that both the FINE and PCA jointly embedded the training and test sets in an unsupervised manner, while the SVM was trained on the embedded space using only the training data.

Fig. 5 illustrates that the embedding calculated with FINE outperforms using PCA as a means of dimension reduction. The classification rates are shown with a one standard deviation confidence interval, and FINE with a dimension as low as $d = 10$ generates results comparable to those of a PCA embedding with $d = 50$. To ease any concerns that LEM is simply a better method for embedding these multinomial PDFs, we calculated an embedding with LEM in which each PDF was viewed as an euclidean vector with the L_2-distance used as a dissimilarity metric. This embedding performed much worse than FINE using the same form of dimension reduction and the same linear kernel SVM, while comparable to PCA only in very low dimensions.

4.2.2 *Supervised FINE*

Allowing FINE to use supervised methods for embedding can improve classification performance. By embedding with CCDR [14], which is essentially LEM with an additional tuning parameter defining the emphasis on class labels in the embedding, we now compare FINE to the diffusion kernels method developed by

[2] http://people.csail.mit.edu/jrennie/20Newsgroups/.

[3] http://www.cs.uiuc.edu/homes/dengcai2/Data/TextData.html.

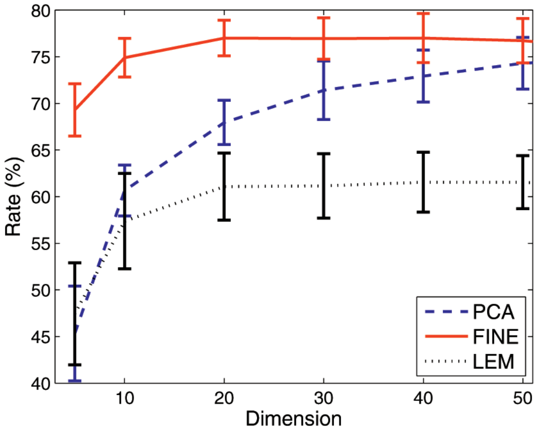


Fig. 5. Classification rates for low-dimensional embedding using different methods for dimension reduction. One standard deviation confidence intervals shown over 20-fold cross validation.

Lafferty and Lebanon [1] for the purpose of document classification. This method uses the full term frequency representation of the data and does not utilize any dimensionality reduction. We stress this difference to determine whether or not using FINE for dimension reduction can generate comparable results.

We now illustrate the classification performance in a "one-versus-all" setting, in which all samples from a single class were given a positive label (i.e., 1) and all remaining samples were labeled negatively (i.e., -1). In the FINE setting, we first subsampled from the recommended training and test indexes, using a test set size of 200, then used CCDR to embed the combined sets into $\mathbb{R}^d$, with $d \in [5, 95]$ chosen to maximize classification performance with a linear kernel SVM. For the diffusion kernels setting, the kernel used was

$$K(X,Y) = (4\pi t)^{-\frac{n}{2}} \exp\left(-\frac{1}{t}\arccos^2\left(\sqrt{X}\cdot\sqrt{Y}\right)\right),$$

where we chose parameter value t which optimized the classification performance at each iteration. The experimental results of performance versus training set size, over a 20-fold cross validation, are shown in Table 1, where the highest performance at each range is emphasized.

Analysis shows that FINE can significantly improve upon the deficiencies of the diffusion kernels method in the low sample size region. By viewing each document as a coarse approximation of the overriding class PDF, it is easy to see that for low sample sizes, the estimate of the within class PDF generated by the diffusion kernels will be highly variable, which leads to poor performance. By reducing the dimension with FINE, the variance is limited to significantly fewer dimensions, yielding better classification performance than using the entire multinomial distribution. As the number of training samples increases, the approximation of the geodesic improves, yielding improved classification performance with FINE. However, the negative effect of dimensionality is also reduced, which allows the diffusion kernels method to better approximate the multinomial PDF representative of each class. This reduction in variance across all dimensions ensures that a few anomalous documents will not have the same drastic effect as they would in the low sample size region. As such, in some instances, the performance gain surpasses that of FINE, due to the fact that the *curse of dimensionality* was alleviated by the increase in sample size. We note, however, that in all cases, FINE performs competitively with a leading document classification method which utilizes the full dimensional data.

TABLE 1
Results on 20 Newsgroups Corpus,
Comparing FINE to the Diffusion Kernel Method

		FINE		Diffusion Kernels	
Task	L	Mean	STD	Mean	STD
comp.*	80	**85.8250**	2.8713	83.0250	3.4469
	200	**87.9750**	2.3978	87.8500	2.2775
	400	**89.8000**	2.0926	89.6250	1.9992
	600	90.6500	2.0970	**91.3000**	2.4677
rec.*	80	**86.3500**	2.0462	82.0000	3.8251
	200	**89.5500**	1.4133	86.8750	2.1143
	400	**91.4750**	2.2152	90.7000	2.0545
	600	92.7500	1.2722	**93.1000**	2.0494
sci.*	80	**80.3750**	3.3280	77.4750	4.2286
	200	**83.4000**	2.9585	82.2000	3.0236
	400	86.1750	2.2021	**86.2000**	2.2325
	600	**87.1750**	2.9212	87.0500	2.9731
talk.*	80	**90.4250**	2.8895	85.9250	3.6859
	200	**92.6500**	1.8503	89.7750	3.1518
	400	**93.1000**	1.9775	92.4750	2.1672
	600	**94.7500**	1.3908	94.3750	1.5634

Performance (classification rate in percent) is reported for different training set sizes L, over a 20-fold cross validation.

5 Conclusions

The assumption that high-dimensional data lie on a Riemannian manifold in euclidean space is based on the ease of implementation due to the wealth of knowledge and methods based on euclidean geometry. This assumption is not viable in many practical problems, as there is often no straightforward and meaningful euclidean representation of the data. In these situations, it is more appropriate to assume that the data are a realization of some PDF which lies on a statistical manifold. Using information geometry, we have shown the ability to find a low-dimensional embedding of the manifold, which allows us to reconstruct it in a low-dimensional euclidean space.

We have illustrated FINE's ability to be used in a variety of learning tasks such as visualization and classification, on a multitude of problems which may seem to have little to nothing in common, such as flow cytometry and document classification. The only commonality between the problems is that each are based around data which have no straightforward euclidean representation, which is the only setting needed to utilize FINE. In future work, we plan to utilize different classification methods (such as k-NN and using different SVM kernels) to maximize our document classification performance. This includes constraining our dimensionality reduction to a sphere, which will allow the use of diffusion kernels in a low-dimensional space. We also plan to study the effect of using out-of-sample extension, rather than jointly embedding the training and test sets. Lastly, we will continue to find applications which benefit from FINE, such as Internet anomaly detection, spam analysis, and object recognition.

Acknowledgments

This work is partially funded by US National Science Foundation, grant No. CCR-0325571.

References

[1] J. Lafferty and G. Lebanon, "Diffusion Kernels on Statistical Manifolds," *J. Machine Learning Research*, vol. 6, pp. 129-163, Jan. 2005.

[2] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, "Face Recognition with Image Sets Using Manifold Density Divergence," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 581-588, June 2005.

[3] S.-M. Lee, A.L. Abbott, and P.A. Araman, "Dimensionality Reduction and Clustering on Statistical Manifolds," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1-7, June 2007.

[4] A. Srivastava, I.H. Jermyn, and S. Joshi, "Riemannian Analysis of Probability Density Functions with Applications in Vision," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1-8, June 2007.

[5] J. Salojarvi, S. Kaski, and J. Sinkkonen, "Discriminative Clustering in Fisher Metrics," *Proc. Int'l Conf. Artificial Neural Networks and Neural Information Processing,* pp. 161-164, June 2003.

[6] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," *Advances in Neural Information Processing Systems,* T.G. Dietterich, S. Becker, and Z. Ghahramani, eds., vol. 14, MIT Press, 2002.

[7] J.B. Tenenbaum, V. de Silva, and J.C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science,* vol. 290, pp. 2319-2323, 2000.

[8] S. Roweis and L. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science,* vol. 290, no. 1, pp. 2323-2326, 2000.

[9] T. Cox and M. Cox, *Multidimensional Scaling.* Chapman & Hall, 1994.

[10] R. Kass and P. Vos, *Geometrical Foundations of Asymptotic Inference.* John Wiley and Sons, 1997.

[11] S. Amari and H. Nagaoka, *Methods of Information Geometry (Translations of Mathematical Monographs),* vol. 191, Am. Math. Soc. and Oxford Univ. Press, 2000.

[12] S.K. Zhou and R. Chellappa, "From Sample Similarity to Ensemble Similarity: Probabilistic Distance Measures in Reproducing Kernel Hilbert Space," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 6, pp. 917-929, June 2006.

[13] K.M. Carter, "Dimensionality Reduction on Statistical Manifolds," PhD thesis, Univ. of Michigan, Jan. 2009.

[14] R. Raich, J.A. Costa, and A.O. Hero, "On Dimensionality Reduction for Classification and Its Application," *Proc. IEEE Int'l Conf. Acoustic Speech and Signal Processing,* vol. 5, May 2006.

[15] K.M. Carter, R. Raich, W.G. Finn, and A.O. Hero, "Information Preserving Component Analysis: Data Projections for Flow Cytometry Analysis," *IEEE J. Selected Topics in Signal Processing,* special issue on digital image processing techniques for oncology, vol. 3, no. 1, pp. 148-158, Feb. 2009.

[16] W.G. Finn, K.M. Carter, R. Raich, and A.O. Hero, "Analysis of Clinical Flow Cytometric Immunophenotyping Data by Clustering on Statistical Manifolds: Treating Flow Cytometry Data as High-Dimensional Objects," *Cytometry Part B: Clinical Cytometry,* vol. 76B, no. 1, pp. 1-7, Jan. 2009.

[17] G. Terrell, "The Maximal Smoothing Principle in Density Estimation," *J. Am. Statistical Assoc.,* vol. 85, no. 410, pp. 470-477, June 1990.

[18] H. Kim, P. Howland, and H. Park, "Dimension Reduction in Text Classification with Support Vector Machines," *J. Machine Learning Research,* vol. 6, pp. 37-53, Jan. 2005.

[19] S. Huang, M.O. Ward, and E.A. Rundensteiner, "Exploration of Dimensionality Reduction for Text Visualization," *Proc. IEEE Third Int'l Conf. Coordinated and Multiple Views in Exploratory Visualization,* pp. 63-74, July 2005.

[20] K.M. Carter, R. Raich, and A.O. Hero, "Fine: Information Embedding for Document Classification," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing,* pp. 1861-1864, Apr. 2008.