FINE: INFORMATION EMBEDDING FOR DOCUMENT CLASSIFICATION

Kevin M. Carter[1]*, *Raviv Raich*[2], *Alfred O. Hero III*[1]

[1] Department of EECS, University of Michigan, Ann Arbor, MI 48109
[2] School of EECS, Oregon State University, Corvallis, OR 97331
{kmcarter,hero}@umich.edu, raich@eecs.oregonstate.edu

ABSTRACT

The problem of document classification considers categorizing or grouping of various document types. Each document can be represented as a bag of words, which has no straightforward Euclidean representation. Relative word counts form the basis for similarity metrics among documents. Endowing the vector of term frequencies with a Euclidean metric has no obvious straightforward justification. A more appropriate assumption commonly used is that the data lies on a statistical manifold, or a manifold of probabilistic generative models. In this paper, we propose calculating a low-dimensional, information based embedding of documents into Euclidean space. One component of our approach motivated by information geometry is the Fisher information distance to define similarities between documents. The other component is the calculation of the Fisher metric over a lower dimensional statistical manifold estimated in a nonparametric fashion from the data. We demonstrate that in the classification task, this information driven embedding outperforms both a standard PCA embedding and other Euclidean embeddings of the term frequency vector.

Index Terms— Manifold learning, Riemannian manifold, geodesics, text classification, information geometry

1. INTRODUCTION

Document classification is an area of machine learning in which it is desired to distinguish between different classes of documents, assuming each document can be represented as a 'bag of words'. Often this task is performed by first using Principal Components Analysis (PCA), which is optimal for Euclidean data, to reduce the dimension of the data and reduce the effect of the *curse of dimensionality*. This type of ad hoc feature vector extraction has been called the "dirty laundry" of machine learning [1]. The problem of document classification is one in which the data has no straightforward Euclidean representation, as each set is a collection of words from a dictionary, which leads to suboptimal processing and information loss. When a statistical model is available, the process of obtaining a feature vector can be done optimally by extracting the model parameters for a given data set. We are interested in extending this approach to the case in which the data follows an unknown parametric statistical model.

A document can be viewed as a realization of some overriding probability distribution, in which different distributions will create different documents. For example, in a newsgroup about computers you could expect to see multiple instances of the term "laptop", while a group discussing recreation may see many occurrences of "sports". The counts of "laptop" in the recreation group, or "sports" in the computer group would predictably be low. As such, the distributions between articles in computers and recreation should be distinct. A standard method for differentiating document classes is to form a probability distribution over a dictionary and use methods of information geometry to determine a similarity between data sets [2]. To the best of our knowledge, most metrics in document classification which are based on word probabilities [3] do not restrict the probability manifold to be a lower dimensional parametric manifold. As a result, a geodesic may go through many probability models that are not admissible in the context of document classification (e.g., a text with only the words 'the' and 'of'). Our approach learns the manifold of admissibility of a probability model from a training set and constructs geodesics based on such manifolds.

In this paper, we utilize the framework presented in [4], which we now refer to as Fisher Information Non-parametric Embedding (FINE), towards the problem of document classification. By viewing each document as a realization of some distribution function lying on a statistical manifold, we are able to generate an information based embedding into a low-dimension Euclidean space. These methods are entirely nonparametric and make no Euclidean assumptions of the data. While previous work has been presented using statistical manifolds for document classification, these methods are geared towards modeling the data [2] and finding optimal classification methods for the data [5]. Our work is restricted to the pre-processing to obtain a low-dimensional representation of the data. We will show using existing classification methods that an information based embedding with FINE outperforms methods optimized for Euclidean data.

*Acknowledgement: This work is partially funded by the National Science Foundation, grant No. CCR-0325571.

2. LEARNING ON STATISTICAL MANIFOLDS

In [4] we presented a framework for learning on statistical manifolds for the purposes of visualization and clustering. We now apply that framework towards the problem of document classification. While details on the theory behind FINE can be found in [4], we shall give a brief overview of the methods.

2.1. Statistical Manifolds

If we consider $\mathcal{M}$ to be a family of probability density functions (PDFs) on the set $\mathcal{X}$, in which each element of $\mathcal{M}$ is a PDF which can be parameterized by $\theta = \left[\theta^1, \ldots, \theta^n\right]$, then $\mathcal{M}$ is known as a statistical model on $\mathcal{X}$. Specifically, let

$$\mathcal{M} = \{p(x \mid \theta) \mid \theta \in \Theta \subseteq \mathbb{R}^n\}, \tag{1}$$

with $p(x \mid \theta)$ satisfying

$$p(x \mid \theta) \geq 0, \, \forall x \in \mathcal{X} \tag{2}$$

$$\int p(x \mid \theta) \, dx = 1.$$

Additionally, there exists a one-to-one mapping between θ and $p(x \mid \theta)$.

We describe only the case for continuum on the set $\mathcal{X}$, however if $\mathcal{X}$ was discrete valued, equation (2) will still apply by switching $\int p(x \mid \theta) \, dx = 1$ with $\sum p(x \mid \theta) = 1$.

When associated with the Fisher information metric, $\mathcal{M}$ is known as a *statistical manifold*. The Fisher information measures the amount of information a random variable X contains in reference to an unknown parameter θ. We define the Fisher information matrix $[\mathcal{I}(\theta)]$, whose elements consist of the Fisher information with respect to specified parameters

$$\mathcal{I}_{ij} = -E\left[\frac{\partial}{\partial \theta^i} \log f(X;\theta) \frac{\partial}{\partial \theta^j} \log f(X;\theta)\right]. \tag{3}$$

For a parametric family of probability distributions, it is possible to define a Riemannian metric using the Fisher information matrix, known as the information metric. The information metric distance, or Fisher information distance, between two distributions $p(x;\theta_1)$ and $p(x;\theta_2)$ is:

$$D_F(\theta_1, \theta_2) = \min_{\substack{\theta(\cdot): \\ \theta(0)=\theta_1 \\ \theta(1)=\theta_2}} \int_0^1 \sqrt{\left(\frac{d\theta}{d\beta}\right)^T \mathcal{I}(\theta) \left(\frac{d\theta}{d\beta}\right)} d\beta. \tag{4}$$

2.1.1. Hellinger Distance

We may approximate the Fisher information distance by the Hellinger distance, which is defined as:

$$D_H(p,q) = \sqrt{\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx},$$

and is related to the information distance in the limit by

$$2D_H(p,q) \to D_F(p,q) \tag{5}$$

as $p \to q$ [6].

While there are other methods of approximating the Fisher information distance (i.e. Kullback-Leibler divergence, Renyi-α entropy), we choose to use the Hellinger distance for the purposes of document classification. Since the PDFs can be considered sparse multinomial distributions, the Hellinger distance avoids any of the divide-by-zero issues associated with other metrics. One should note that when dealing with multinomial distributions, the approximation

$$D_C(p,q) = 2 \arccos \int \sqrt{p \cdot q} \to D_F(p,q),$$

as this is the natural metric on the sphere [6]. D_H and D_C are related by a monotonic transformation function, so we restrict our metric to that of the Hellinger distance.

2.2. Approximation of Distance on Statistical Manifolds

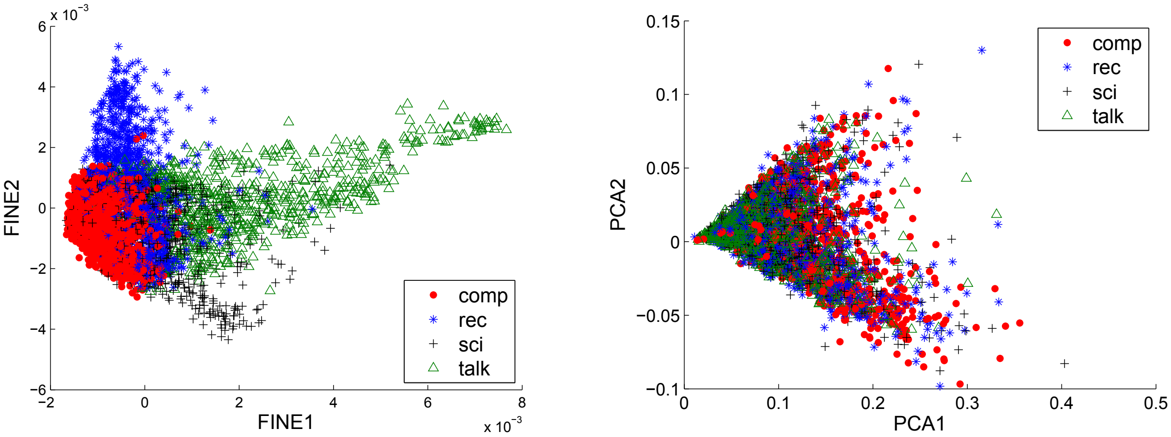
As noted earlier (5), $2D_H(p_1,p_2) \to D_F(p_1,p_2)$ as $p_1 \to p_2$. If p_1 and p_2 do not lie closely together on the manifold, the Hellinger distance becomes a weak approximation of the Fisher information distance. However, a good approximation can still be achieved if the manifold is densely sampled between the two end points by defining the path between p_1 and p_2 as a series of connected segments, and summing the length of those segments. Specifically, given the set of n probability density functions $\mathcal{P} = \{p_1, \ldots, p_n\}$, the Fisher information distance between p_1 and p_2 can be estimated as:

$$D_F(p_1,p_2) \approx \min_{m, \left\{p_{(1)}, \ldots, p_{(m)}\right\}} \sum_{i=1}^{m} 2D_H(p_{(i)}, p_{(i+1)}),$$

where $p_{(1)} = p_1$ and $p_{(m)} = p_2$. Intuitively, this estimate calculates the length of the shortest path between points in a connected graph on the well sampled manifold.

2.3. Dimensionality Reduction

Given a matrix of dissimilarities between entities, many algorithms have been developed to find a low dimensional embedding of the original data $\psi : \mathcal{M} \to \mathbb{R}^d$. These techniques have been classified as a group of methods referred to as Multi-Dimensional Scaling (MDS). There are supervised methods, which are generally used for classification purposes, and unsupervised methods, which are often used for clustering and manifold learning. For our purposes, we will use the unsupervised method Laplacian Eigenmaps [7], which will reveal any natural separation or clustering of the data sets. This allows us to find a single low-dimensional coordinate representation of each high-dimensional, large sample, data set. Once this



(a) FINE (b) PCA

Fig. 1. 2-dimensional embeddings of 20Newsgroups data. The data displays some natural clustering, in the information based embedding, while the PCA embedding does not distinguish between classes.

```
Algorithm 1 Fisher Information Non-parametric Embedding
Input: Collection of PDFs $\mathcal{P} = \{p_1, p_2, \ldots, p_N\}$ and the
    desired embedding dimension $d$
  1: Calculate $D$, where $D(i,j) = \hat{D}_F(p_i, p_j)$
  2: $\boldsymbol{Y} = \text{embed}(D, d)$
Output: $d$-dimensional embedding of $\mathcal{P}$, into Euclidean
    space $\boldsymbol{Y} \in \mathbb{R}^{d \times N}$
```

Euclidean embedding is obtained, we can use learning methods such as Support Vector Machines (SVMs) to perform our classification task.

2.4. Algorithm

Algorithm 1, which we refer to as the Fisher Information Non-parametric Embedding (FINE), illustrates the ability to find a low-dimensional embedding of a collection of PDFs. If we assume each data set is a realization of an underlying probability density, and each of those densities lie on a manifold with some natural parameterization, then this embedding can be viewed as an embedding of the actual manifold into Euclidean space. Note that in line 2, 'embed(D, d)' refers to using any multi-dimensional scaling method (such as Laplacian Eigenmaps) to embed the approximation of the Fisher information distance matrix into $\mathbb{R}^d$.

3. SIMULATION RESULTS

The problem of document classification is an interesting application for FINE. Given a collection of documents of known class, we wish to best classify a document of unknown class. 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

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

Given these term frequency representations, the problem of document classification fits directly into our framework.

For illustration, we will utilize the well known 20 Newsgroups data set[1], which is commonly used for testing document classification methods. This set contains word counts for postings on 20 separate newsgroups. We choose to restrict our simulation to the 4 domains with the largest number of sub-domains (comp.*, rec.*, sci.*, and talk.*), and wish to 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). First, we utilize unsupervised methods to see if the natural geometry exists between domains. We note that the data was preprocessed to remove all words that occur in 5 or less documents[2]. Using Laplacian Eigenmaps on the dissimilarities calculated with the Hellinger distance, we found an embedding $\mathcal{P} \rightarrow \mathbb{R}^2$. Figure 1(a) shows the natural geometric separation between the different document classes, although there is some expected overlap. Contrarily, a Principal Components Analysis (PCA) embedding (Fig. 1(b)) does not demonstrate the same natural clustering. 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

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

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

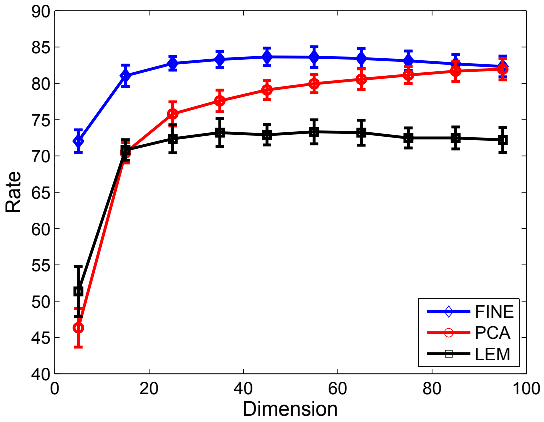


Fig. 2. Classification rates for low-dimensional embedding using different methods for dimensionality reduction

of the 20 Newsgroups set does not exhibit any natural class separation due to the non-Euclidean nature of the data.

We now compare the classification performance of FINE to that of PCA. In the case of document classification, dimensionality reduction is important as the natural dimension (i.e. number of words) for the 20 Newsgroups data set is 26,214. We test performance for low dimensional embeddings $\mathcal{P} \rightarrow \mathbb{R}^d$ for $d = 5$ to $d = 95$. Following each embedding, we apply an SVM with a linear kernel to classify the data in an 'all-vs-all' setting (i.e. classify each test sample as one of 4 different potential classes). The training and test sets were separated according to the recommended indices, and each set was randomly sub-sampled for computational purposes (2413 training samples, 1607 test samples). Both the FINE and PCA settings jointly embed the training and test sets.

Figure 2 illustrates that the embedding calculated with FINE outperforms using PCA as a means of dimensionality reduction. The classification rates are shown with a 1-standard deviation confidence interval, and FINE with a dimension as low as $d = 25$ generates results comparable to those of a PCA embedding with $d = 95$. To ease any concerns that Laplacian Eigenmaps (LEM) is simply a better method for embedding these multinomial PDFs, we calculated an embedding with LEM in which each PDF was viewed as a Euclidean vector with the L_2-distance used as a dissimilarity metric. This form of embedding performed much worse than the information based embedding using the same form of dimensionality reduction and the same linear kernel SVM, while comparable to the PCA embedding in very low dimensions.

4. CONCLUSIONS

We have presented the ability to derive an information based embedding for a data set consisting of term frequency representations of documents. By not making Euclidean assumptions on the data, we find a low dimensional representation which yields high classification rates with a linear SVM, outperforming an embedding calculated with Principal Component Analysis. Although PCA is optimal for Euclidean data, it is clear that there is no straightforward way to represent a document in Euclidean space. The standard term frequency representation is clearly non-Euclidean and is better viewed as a probability distribution. As such, using information geometry to learn on the underlying statistical manifold is more appropriate than the ad hoc manner of Euclidean feature vector extraction.

In future work we intend to test our FINE algorithm with various different manifold learning methods. While we currently choose to use Laplacian Eigenmaps to generate our low dimensional representation, we will look into other multidimensional scaling methods to determine which gives the best performance. In this paper we focused on unsupervised methods in order to garner a fair comparison to PCA, however we plan to utilize supervised methods of dimensionality reduction to see if we can generate better classification performance than an SVM (with a linear or diffusion kernel [5]) on the full dimensional data set.

5. REFERENCES

[1] T. Dietterich, "Ai seminar," Carnegie Mellon, 2002.

[2] G. Lebanon, "Information geometry, the embedding principle, and document classification," in *Proceedings of the 2nd International Symposium on Information Geometry and its Applications*, 2005.

[3] G. Lebanon, "Metric learning for text documents," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 497–508, April 2006.

[4] K. M. Carter, R. Raich, and A. O. Hero, "Learning on statistical manifolds for clustering and visualization," in *Proceedings of Forty-Fifth Annual Allerton Conference on Communication, Control, and Computing*, September 2007, to appear. http://tbayes.eecs.umich.edu/kmcarter/LearnStatMan.html.

[5] J. Lafferty and G. Lebanon, "Diffusion kernels on statistical manifolds," *Journal of Machine Learning Research*, vol. 6, pp. 129–163, Jan 2005.

[6] R. Kass and P. Vos, *Geometrical Foundations of Asymptotic Inference*, Wiley Series in Probability and Statistics. John Wiley and Sons, NY, USA, 1997.

[7] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems, Volume 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2002.