Find the Intrinsic Space for Multiclass Classification *

Meizhu Liu
Department of Computer and Information Science and Engineering
University of Florida
Gainesville, FL, 32611
mliu@cise.ufl.edu

Kefei Liu
Department of Electronic Engineering
City University of Hong Kong
Tat Chee Avenue, Kowloon
Tong, Kowloon, Hong Kong
kefeilau@gmail.com

Xiaojing Ye
School of Mathematics
Georgia Tech
686 Cherry Street
Atlanta, GA, 30332
xye33@math.gatech.edu

ABSTRACT

Multiclass classification is one of the core problems in many applications. High classification accuracy is fundamental to be accepted as a valuable or even indispensable tool in the work flow. In the classification problem, each sample is usually represented as a vector of features. Most of the cases, some features are usually redundant or misleading, and high dimension is not necessary. Therefore, it is important to find the intrinsically lower dimensional space to get the most representative features that contain the best information for classification. In this paper, we propose a novel dimension reduction method for multiclass classification. Using the constraint of the triplet set, our proposed method projects the original high dimensional feature space to a much lower dimensional feature space. This method enables faster computation, reduce the space needed, and mostly importantly produces more meaningful representations that leads to better classification accuracy.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Dimension reduction, multiclass classification, triplet

1. INTRODUCTION

ISABEL '11, October 26-29, Barcelona, Spain

Classification is an important task in computer vision, machine learning, and pattern recognition etc. Nonparametric classification techniques, such as nearest neighbor voting or template matching [20, 14], can be flexible and powerful representations for joint classification, clustering and retrieval. But both nearest neighbor voting and template matching are sensitive to high dimensional feature space, due to the reason that high dimension is not necessary or some features may be misleading. Therefore, it becomes important to find the lower dimensional space which is able to represent the data set intrinsically. Dimension reduction is such a method.

Dimension reduction is an important step to improve the classification accuracy and decrease computational and spacial complexity. There are lots of dimension reduction techniques in the literature [2, 12]. Dimension reduction is usually achieved by using feature selection [15] or feature projection. Feature projections can be done in different ways: minimizing the reconstruction error as principal component analysis (PCA) [5, 10]; preserving distances in the original space, e.g. multidimensional scaling (MDS) [4], ISOMAP [19] which uses geodesic distances in the data space, diffusion maps which uses diffusion distances in the data space, and curvilinear component analysis [11]; maximizing class-data separation as linear discriminant analysis (LDA) [5]; retaining the linear relationship between locality neighbors, e.g., neighborhood component analysis (NCA) [8], locally linear embedding (LLE) [17]; preserving all pairwise distances between nearest neighbors (in the inner product space), while maximizing the distances between points that are not nearest neighbors, e.g., maximum variance unfolding (MVU) [21]. We follow the principle that keeps the locality of data belonging to the same class closer and maps data belonging to different classes further, in the graph-induced subspace, which is similar to Laplacian Eigenmap [1] and Locality Preserving Projection [9]. We will propose a novel triplet constraint based method to reduce the dimension of the feature space, which enables accessible and accurate multiclass classification accuracy.

In this paper, we exploit a novel triplet based graph embedding (TGE) to project data into an even lower dimensional subspace.A triplet is composed of three data samples, representing the approximation relationship between them. This will be explained at length later. TGE makes the data samples from the same class getting closer and samples from different classes moving away, to make nearest neighbor voting, template matching or any other classification methods

more robust and semantically interpretable. Therefore, it greatly enhances the classification accuracy.

The rest of the paper is organized as follows. In section 2, we present our proposed dimension reduction method, with triplets as the constraints, and explain that in detail. This is unlike all the existing dimension reduction techniques. In section 3, we evaluate our dimension reduction technique on various benchmark data sets, which belong to the challenging UCI machine learning repository [7]. Finally, this paper is concluded with discussion in section 4.

2. PROPOSED METHOD: TRIPLET BASED GRAPH EMBEDDING

The input for our proposed dimension reduction includes a set of N points $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \subset \mathbb{R}^n$, along with their labels $\mathcal{X} = \{l_1, l_2, \cdots, l_N\} \subset \mathbb{C}$. $\mathbb{C}$ is the label set, which is usually a set composed of integers, e.g., for binary classification problems, $\mathbb{C}$ is usually $\{1, -1\}$.

The goal of dimension reduction is to give small distances between instances to be matched and large distance for others, in the reduced dimensional space. There are a number of ways to design dimension reduction. One popular technique is to perform dimension reduction according to the relations of the training samples. One type of relation is equivalence constrained, where equivalence constraints are provided for pairs $(\mathbf{x_i}, \mathbf{x_j})$, each associated with a binary label of "similar" or "dissimilar" [18]. Another relation representation often used in information retrieval is the proximity relationships [13] over triplet set $\mathcal{T} = \{(i, j, k)\}$, meaning that $\mathbf{x_i}$ is closer to $\mathbf{x_j}$ than to $\mathbf{x_k}$. Here $\mathbf{x_i}$ is the feature vector representation for the training sample i.

The goal of dimension reduction is to learn a function $f : \mathcal{X} \mapsto \mathcal{Y}$, where $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N\} \subset \mathbb{R}^{\tilde{n}}$, and $\tilde{n} \ll n$, such that $d(\mathbf{x_i}, \mathbf{x_j}) < d(\mathbf{x_i}, \mathbf{x_k})$, and the distance between two samples $\mathbf{x_i}$ and $\mathbf{x_j}$ in the reduced dimensional space is

$$d(\mathbf{x_i}, \mathbf{x_j}) = (f(\mathbf{x_i}) - f(\mathbf{x_j}))^T (f(\mathbf{x_i}) - f(\mathbf{x_j}))\,, \qquad (1)$$

where ^{T} is the vector/matrix transpose transformation. The optimal f should maximize the following function

$$E = \sum_{ijk} w(i, j, k)(d(\mathbf{x_i}, \mathbf{x_k}) - d(\mathbf{x_i}, \mathbf{x_j})), \qquad (2)$$

under some appropriate constraints. $w(i, j, k)$ is the weight for the triplet (i, j, k), and usually it is set to be the uniform distribution, i.e., $w(i, j, k) = 1/|\mathcal{T}|$ where $|\mathcal{T}|$ is the cardinality of $\mathcal{T}$. This objective function ensures $\mathbf{y}_i$ and $\mathbf{y}_j$ to be close if $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same class, and vise versa.

Triplet based graph embedding (TGE) is a comprehensive strategy to simultaneously maximize the similarity between data pairs of the same class and minimize the similarity between two points rooted from different classes. In other words, we optimize on mapping the same class data to proximity subspaces, while projecting different class data samples to be far apart, explicitly. Once satisfying this, we can classify the testing samples in the reduced dimensional space.

2.1 Choose the mapping function

Various choices of the mapping function f have been proposed recently, e.g. linear mapping, kernel mapping and tensor mapping. We use linear mapping because of its simplicity and generality. A linear mapping function f is described as

$$\begin{aligned} &\mathbf{y} = f(\mathbf{x}) = \mathbf{P}^T\mathbf{x}, \\ \text{subject to: } &\|\mathbf{P}\|_F = 1, \\ &\mathbf{P} \in \mathbb{R}^{n\times\tilde{n}}\,, \tilde{n} \ll n\,, \end{aligned} \qquad (3)$$

where $\|\cdot\|_F$ is the Frobenius norm, and the constraint $\|M\|_F = 1$ removes the scaling effect.

Plugging (3) into (1), we get

$$\begin{aligned} d(\mathbf{x_i}, \mathbf{x_j}) &= \|\mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j\|^2 \\ &= (\mathbf{x_i} - \mathbf{x_j})^T\mathbf{P}\mathbf{P}^T(\mathbf{x_i} - \mathbf{x_j})\,. \end{aligned} \qquad (4)$$

Plugging (1) into (2), we get

$$\begin{aligned} &E = tr\left(\mathbf{P}\sum_{(i,j,k)} w(i, j, k)(\|\mathbf{x}_i - \mathbf{x}_k\|^2 - \|\mathbf{x}_i - \mathbf{x}_j\|^2)\mathbf{P}^T\right), \\ &\text{subject to: } \|\mathbf{P}\|_F = 1\,. \end{aligned} \qquad (5)$$

Eq. (5) can be solved very quickly using gradient descent technique along with iterative projections [16].

2.2 Choose the dimension of the reduced space

The dimension of the reduced space $\tilde{n}$ can be determined in many ways, the most popular ways are:

1. The variance is covered to some extent, e.g. 90% percent.
2. Some small numbers like 1, 2 or 3 for visualization.
3. Chosen to be the number of positive eigenvalues of the covariance matrix of the data set.
4. A fixed number according to the needs of customers.
5. Chosen to be the dimension which gives optimal classification performance on the training/validation data set.
6. Chosen to be the number such that the loss function (5) is minimized.

We will use the last one in the applications of this paper.

2.3 Evaluate the dimension reduction methods

The effectiveness of dimension reduction can be evaluated according to several criteria, such as information gain [3], and Fisher score [6]. We validate the effectiveness of our proposed dimension reduction technique using Fisher Score, where the class separability between different classes is measured via Fisher's linear discriminant [6].

Let the covariance matrices of the negatives and positives be Σ_- and Σ_+, and the means of the negatives and positives

be μ_- and μ_+, then the Fisher linear discriminant of the binary classes is

$$s = (\mu_+ - \mu_-)^T (\mathbf{\Sigma}_+ + \mathbf{\Sigma}_-)^{-1} (\mu_+ - \mu_-)\,, \qquad (6)$$

where the larger s is, the more statistically distinguishable negative-positive class distributions will be.

3. EXPERIMENTAL RESULTS

We evaluated our algorithm on numerous public domain data sets from the UCI machine learning repository [7]. The UCI repository [7] is a collection of popular databases that have been extensively used for analyzing machine learning especially classification techniques. The repository contains very noisy data (e.g. waveform) as well as relatively clean data, which is optimal for testing classification algorithms. We selected many data sets from the UCI repository. The selected data sets include noisy and relative clean data sets, cover small size to large size data sets in terms of number of instances in the data sets, and range from low dimension to high dimension in terms of number of attributes per sample of the data sets. The description of the selected data sets is shown in Table 1. We compared our classification results with those generated from MRMR feature selection.

We form the triplets in the following way. For each training sample $\mathbf{x_i}$, we find all the training samples $\{\mathbf{x_j}\}_{j=1}^{n}$ that belong to the same class as $\mathbf{x_i}$, and all other training samples $\{x_k\}_{k=1}^{m}$ which belong to different classes from the class of $\mathbf{x_i}$. Then (i, j, k) will form a triplet, requiring $d(\mathbf{x_i}, \mathbf{x_j}) < d(\mathbf{x_i}, \mathbf{x_k})$. We repeat the same process on each training sample to build more triplets, in a similar way. All the triplets form a triplet set $\mathcal{T}$, which will be used as inputs for dimension reduction algorithm to optimize the matrix M.

We use 5-fold cross-validation, i.e. 80% of the samples are for training and validation, and 20% of the samples are for testing. We determine the parameters for TGE and MRMR algorithms during the training and validation period, and the parameters are set to be those maximizing the classification accuracy of the training samples.

TGE is capable to increase the discriminant between different classes in the projected feature subspaces, both visually and numerically. This is validated on the data sets of Connectionist Bench (Sonar, Mines vs. Rocks), Statlog (German Credit Data), and UJI Pen Characters. For comparison, we plot the first MRMR selected original features and the first projected dimensions after TGE, on testing data sets in Fig. 1. The Fisher linear discriminant score for the first three MRMR selected features on the Connectionist Bench (Sonar, Mines vs. Rocks) data set is 0.3215, but after TGE, the score increases to 0.8910. For the Statlog (German Credit Data) data set, the score increases from 0.1137 to 0.5902, reflecting the impact of TGE. For the UJI Pen Characters data set, the score increases from 0.2608 to 0.8351. The numerical results demonstrate that our TGE technique indeed enlarges the class separability between different classes.

The experimental results of classification on the testing data sets are shown in Table 2. The results show that TGE is capable to give a much higher classification accuracy.

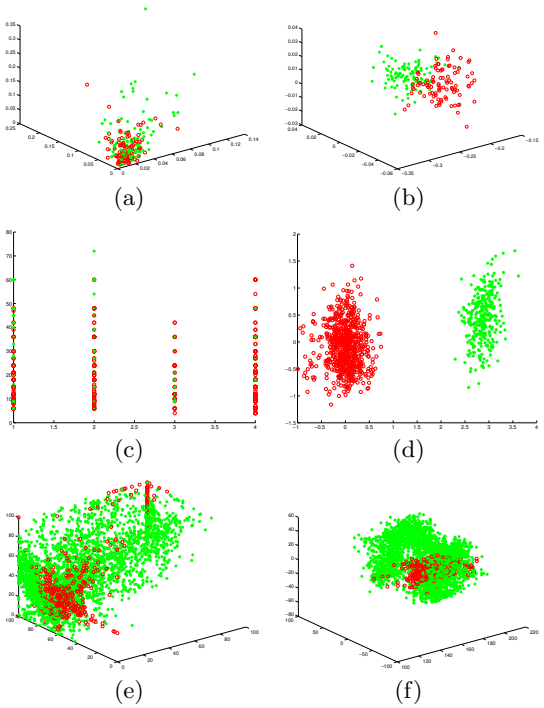

Figure 1: Plot of the data samples (testing) according to the first three features selected by MRMR (a) and the first three dimensions after dimension reduction (b) on the Connectionist Bench (Sonar, Mines vs. Rocks) data set. Plot of the data samples (testing) according to the first two features selected by MRMR (c) and the first two dimensions after dimension reduction (d) on the German Credit data set. Plot of the data samples (testing) according to the first three features selected by MRMR (e) and the first three dimensions after dimension reduction (f) on the UJI Pen Characters data set. Note that the dimension coordinates on the figures are not directly comparable.

4. CONCLUSIONS

Our main contribution is summarized as follows. We presented a new dimension reduction method, namely triplet based graph embedding (TGE). This method is triplet constrained, where a triplet represents the proximity relationships between samples. TGE can be applied to classify multiclass data sets. We evaluated this method on a number of data sets from the benchmark UCI machine learning repository. The results show that our method reduces or eliminates the redundancy between the components of high-dimensional vector data, obtains a compact as well as accurate representation, and enables higher classification accuracy.

5. ACKNOWLEDGMENTS

This work was in part supported by the University of Florida Alumni Fellowship to Meizhu Liu, the City University of Hong Kong Postgraduate Studentship to Kefei Liu, and Georgia Institute of Technology postdoc fellowship to Xiaojing

data set	♯ instances	♯ attributes	description
Breast Cancer Wisconsin (Original)	699	9	breast cancer diagnosis
Pima Indians Diabetes	768	8	diabetes diagnosis
Statlog (German Credit Data)	1000	24	good/bad credit
Heart Disease	303	74	diagnosis of heart disease
Ionosphere	351	33	radar returns from the ionosphere
Liver Disorders	345	6	liver disorders arise from alcohol consumption
Connectionist Bench (Sonar, Mines vs. Rocks)	208	60	sonar signals
UJI Pen Characters	1364	16	handwritten digits
Pima Indians Diabetes	768	8	signs of diabetes

Table 1: Description of the UCI data sets that we use.

data set	MRMR feature selection	Dimension reduction
Breast Cancer Wisconsin (Original)	0.658 ± 0.049	0.701 ± 0.037
Pima Indians Diabetes	0.647 ± 0.051	0.705 ± 0.028
Heart Disease	0.708 ± 0.042	0.801 ± 0.030
Ionosphere	0.670 ± 0.036	0.743 ± 0.021
Liver Disorders	0.702 ± 0.061	0.766 ± 0.024

Table 2: Classification accuracy (mean $\pm$ deviation) for MRMR method and our dimension reduction method.

Ye.

6. REFERENCES

[1] M. Belkin and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, pages 1373–1396, 2003.

[2] J. Bi, K. P. Bennett, M. Embrechts, C. M. Breneman, and M. Song. Dimensionality Reduction via Sparse Support Vector Machines. *Journal of Machine Learning Research*, pages 633–42, 2003.

[3] T. M. Cover and J. A. Thomas. Elements of Information Theory. *Wiley-Interscience, 2nd edition*, 2006.

[4] M. Cox and T. Cox. *Multidimensional Scaling*, pages 315–347. Springer Handbooks Comp. Statistics, 2008.

[5] R. Duda, P. Hart, and D. Stork. Pattern Classification. *Wiley-Interscience, Hoboken, NJ, 2nd edition*, 2000.

[6] R. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7:179–188, 1936.

[7] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[8] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighborhood Component Analysis. *In Advances in Neural Information Processing Systems*, 2004.

[9] X. He and P. Niyogi. Locality Preserving Projections. *In Advances in Neural Information Processing Systems*, 2003.

[10] I. Jolliffe. Principal Component Analysis. *Springer-Verlag*, 1986.

[11] J. A. Lee, A. Lendasse, N. Donckers, and M. Verleysen. A Robust Nonlinear Projection Method. *Proceedings - European Symposium on Artificial Neural Networks*, pages 13–20, 2000.

[12] M. Liu, L. Lu, X. Ye, and S. Yu. Coarse-to-fine Classification using Parametric and Nonparametric Models for Computer-Aided Diagnosis. *20th ACM Conference on Information and Knowledge Management (CIKM)*, 2011.

[13] M. Liu, L. Lu, X. Ye, S. Yu, and M. Salganicoff. Sparse Classification for Computer Aided Diagnosis Using Learned Dictionaries. *14th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 6893:41–48, 2011.

[14] M. Liu, B. vemuri, S. Amari, and F. Nielsen. Total Bregman Divergence and its Applications to Shape Retrieval. *IEEE Computer Vision and Pattern Recognition*, 2010.

[15] H. Peng, F. Long, and C. Ding. Feature Selection Based on Mutual Information: Criteria of Max-dependency, Max-relevance, and Min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1226–1238, 2005.

[16] R. Rockafellar. Convex Analysis. *Princeton University Press*, 1970.

[17] S. Roweis and L. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290:2323–2326, 2000.

[18] C. Shen, J. Kim, L. Wang, and A. van den Hengel. Positive semidefinite metric learning with boosting. *In Proc. Adv. Neural Inf. Process. Syst.*, 2009.

[19] J. Tenenbaum, V. Silva, and J. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290:2319–2323, 2000.

[20] B. Vemuri, M. Liu, S. Amari, and F. Nielsen. Total Bregman Divergence and its Applications to DTI Analysis. *IEEE Transactions on Medical Imaging*, 30(2):475–483, 2011.

[21] K. Q. Weinberger and L. K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision, In Special Issue: Computer Vision and*

Pattern Recognition-CVPR 2005 Guest Editor(s):
Aaron Bobick, Rama Chellappa, Larry Davis,
70(1):77–90, 2005.