Ziqiang WANG[1,2], Xia SUN[1], Xu QIAN[2]

Henan University of Technology (1), China University of Mining & Technology(Beijing) (2)

Efficient Kernel Discriminative Geometry Preserving Projection for Document Classification

Abstract. *A new dimensionality reduction algorithm called kernel discriminative geometry preserving projection (KDGPP) is proposed to cope with document classification. By considering both intraclass geometry and interclass discrimination, KDGPP can not only nonlinearly project documents into lower-dimensional feature space via manifold adaptive kernel function but also reduce the computational complexity with Nyström method. Experimental results demonstrate that KDGPP outperforms other related algorithms in terms of effectiveness and efficiency.*

Streszczenie. *Zaproponowano nowy algorytm do klasyfikacji dokumentów nazwany KDGPP – kernel discriminative geometry preserving projection. Algorytm redukuje złożoność obliczeń numerycznych. (**Algorytm KDGPP do identyfikacji I klasyfikacji dokumentów**)*

Keywords: document classification, dimensionality reduction, kernel discriminative geometry preserving projection(KDGPP).
Słowa kluczowe: klasyfikacja dokumentów, rozpoznawanie znaków.

Introduction

Document classification is a key component for many practical applications such as digital library, opinion analysis, and Web search engine. Usually, document classification is very difficult because of the high dimensionality of documents to be classified. In order to reduce the dimensionality in analyzing documents, many researchers studied to use dimensionality reduction for document classification.

The most popular dimensionality reduction algorithms for document data is latent semantic indexing(LSI) [1], it aims to find the best subspace approximation to the original document space in the sense of minimizing the global reconstruction error. Recently, nonnegative matrix factorization (NMF) [2] is also applied to reduce the dimension of document data. However, NMF only focuses on the global geometrical structure of document data. Thus, both LSI and NMF are not suitable for document classification problems since they do not use valuable class label information while reducing the dimension of document data. Unlike LSI and NMF which are unsupervised, LDA[3] is supervised. LDA aims to find an optimal transformation that maps the data into a lower-dimensional space that minimizes the within-class scatter and simultaneously maximizes the between- class scatter. However, the above dimensionality reduction algorithms see only the global Euclidean structure and cannot discover the nonlinear manifold structure hidden in the high-dimensional data. Therefore, they might not be optimal in discriminating documents with different semantic which is the ultimate goal of document classification.

Recently, a number of manifold learning-based dimensionality reduction algorithm have been developed. One of the key ideas in manifold learning approaches is the so called locally invariant idea, i.e., the nearby points are likely to have the similar embedding [4]. However, these manifold learning algorithms are imperfect for classification tasks, because they only consider the intraclass geometry, while ignore the interclass separability. Marginal Fisher analysis (MFA)[5] is a solution which takes both the intraclass geometry and the interclass discrimination into account. However, MFA fails to consider the discriminative information contained in the nonmarginal samples and has the singularity problem. The discriminative geometry preserving projection(DGPP) [6] method is recently proposed to models both the intraclass geometry and interclass discrimination and never meets the undersampled problem. However, DGPP is a linear dimensionality reduction method, so it cannot well describe complex nonlinear variations of documents with different contents.

In this paper, to enhance the classification performance of DGPP, we propose a new dimensionality reduction algorithm termed kernel DGPP (KDGPP) which is fundamentally based on DGPP and kernel trick. It can precisely models both the intraclass geometry and interclass discrimination, cope with nonlinear reduction of documents with kernel function, and greatly reduce the computational burdens in computing the large scale kernel matrix.

Brief review of DGPP

Discriminative geometry preserving projection(DGPP)[6] is a recently proposed manifold learning algorithm for linear dimensionality reduction, it can implement the discrimination preservation and the local geometry preservation abilities by using average weighed adjacency graph and local linear reconstruction error.

Given a set of documents $\{x_1, x_2, \cdots, x_n\} \subset \mathbb{R}^H$,let $X = [x_1, x_2, \cdots, x_n]$ .The DGPP can be obtained by solving the following maximization problem:

$$\begin{aligned} U_{opt} &= \arg\max_U \sum_{i,j}^{n} h_{ij} \left\| U^T x_i - U^T x_j \right\|^2 - \sum_{i=1}^{n} \left\| U^T x_i - \sum_{j:c_j=c_i} w_{ij} U^T x_j \right\|^2 \\ &= \arg\max_U \operatorname{Tr} \left(U^T X \begin{pmatrix} \left(D - H^T\right)\left(D - H^T\right)^T \\ -\left(I - W^T\right)\left(I - W^T\right)^T \end{pmatrix} X^T U \right) \end{aligned} \tag{1}$$

with the constraint

$$U^T U = I \tag{2}$$

where the weighing factor h_{ij} denotes both the distance weighing information and the class label information in terms of

$$h_{ij} = \begin{cases} \exp\left(-\left\|x_i - x_j\right\|^2 / \sigma^2\right)\left(1/n - 1/n_l\right), & \text{if } c_i = c_j = l \\ 1/n, & \text{if } c_i = c_j \end{cases} \tag{3}$$

where the document class label $c_i \in \{1, 2, \cdots, c\}$ and the lth class contains n_l samples satisfying $\sum_{l=1}^{c} n_l = n$.

In addition, w_{ij} denotes the reconstruction coefficient of x_i which can be linearly reconstructed from the samples x_j with the same class label $c_i = c_j$. By imposing $\sum_{j:c_j=c_i} w_{ij} = 1$ and $w_{ij} = 0$ for $c_i \neq c_j$, w_{ij} can be obtained by solving the following reconstruction error minimization problem:

$$w_{ij} = \arg\min_{w_{ij}} \sum_{i=1}^{n} \|\varepsilon_i\|^2 = \arg\min_{w_{ij}} \sum_{i=1}^{n} \|x_i - w_{ij} x_j\|^2 \tag{4}$$

Thus, we can easily have $w_i = \sum_p C_{i,p}^{-1} / \sum_{p,q} C_{p,q}^{-1}$,wherein $C_{p,q} = (x_i - x_p)^T (x_i - x_q)$ is the local Gram matrix and $c_p = c_q = c_i$.

As can be seen from the above statement, DGPP aims to look for a linear transformation matrix U such that the distances between interclass samples are as large as possible while distances between intraclass samples are as small as possible; and the local geometry of intraclass samples is preserved as much as possible by keeping linear reconstruction error minimization.

Finally, the transformation matrix U are the eigenvectors associated with the largest eigenvalues of the following stand eigenproblem:

$$X \begin{pmatrix} (D-H^T)(D-H^T)^T \\ -(I-W^T)(I-W^T)^T \end{pmatrix} X^T U = \lambda U \tag{5}$$

Thus, the U can also be regarded as the eigenvectors of the matrix $X\left((D-H^T)(D-H^T)^T - (I-W^T)(I-W^T)^T\right)X^T$ associated with the largest eigenvalues. Because DGPP does not compute any matrix inverse for dimensionality reduction, it successfully avoids the singularity problem.

Kernel DGPP (KDGPP)

Although DGPP has attained reasonably good performance in pattern classification, it is a linear dimensionality reduction method and fails to describe complex nonlinear variations of documents. To deal with this limitation, the nonlinear extension of DGPP through kernel trick is introduced in this paper. In the following, we discuss how to perform DGPP in Reproducing Kernel Hilbert Space (RKHS), which gives rise to kernel DGPP (KDGPP) algorithm for nonlinear dimensionality reduction.

To extend the above DGPP to the nonlinear mapping case, we consider the problem in a feature space F induced by some nonlinear mapping

$$\varphi : x \in R^N \to \varphi(x) \in F \tag{6}$$

For a proper chosen φ , an inner product $\langle , \rangle$ can be defined on F , which makes for a so-called reproducing kernel Hilbert space (RKHS). In implementation, the implicit feature vector $\varphi(x)$ does not need to be computed explicitly, while it is just done by computing the inner product of two vectors in F with a kernel function.More specifically,

$$\langle \varphi(x), \varphi(y) \rangle = k(x, y) \tag{7}$$

holds where $k(,)$ is a positive semi-definite kernel function.

Let φ denote the document data matrix in RKHS:

$$\varphi = \varphi(X) = [\varphi(x_1), \varphi(x_2), \cdots, \varphi(x_n)] \tag{8}$$

Then, the eigenvector problem of (5) in RKHS can be written as follows:

$$\varphi(X) \begin{pmatrix} (D-H^T)(D-H^T)^T \\ -(I-W^T)(I-W^T)^T \end{pmatrix} \varphi(X)^T U = \lambda U \tag{9}$$

Because the eigenvector of (9) must lie in the span of all the samples in F , there exist coefficients α_i , $i = 1, 2, \cdots, n$, such that

$$U = \sum_{i=1}^{n} \alpha_i \varphi(x_i) = \varphi\alpha \tag{10}$$

where $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)^T$.

Following some algebraic formulations, we get:

$$\begin{aligned} & \varphi(X) \begin{pmatrix} (D-H^T)(D-H^T)^T \\ -(I-W^T)(I-W^T)^T \end{pmatrix} \varphi(X)^T U = \lambda U \\ \Rightarrow & \varphi \begin{pmatrix} (D-H^T)(D-H^T)^T \\ -(I-W^T)(I-W^T)^T \end{pmatrix} \varphi^T U = \lambda U \\ \Rightarrow & \varphi \begin{pmatrix} (D-H^T)(D-H^T)^T \\ -(I-W^T)(I-W^T)^T \end{pmatrix} \varphi^T \varphi\alpha = \lambda\varphi\alpha \\ \Rightarrow & \varphi^T\varphi \begin{pmatrix} (D-H^T)(D-H^T)^T \\ -(I-W^T)(I-W^T)^T \end{pmatrix} \varphi^T \varphi\alpha = \lambda\varphi^T\varphi\alpha \\ \Rightarrow & K \begin{pmatrix} (D-H^T)(D-H^T)^T \\ -(I-W^T)(I-W^T)^T \end{pmatrix} K\alpha = \lambda K\alpha \\ \Rightarrow & \begin{pmatrix} (D-H^T)(D-H^T)^T \\ -(I-W^T)(I-W^T)^T \end{pmatrix} K\alpha = \lambda\alpha \end{aligned} \tag{11}$$

where K is the kernel matrix, $K_{ij} = k(x_i, x_j)$.

From the derivation process of (11), we can observe that the coefficient vector α are the eigenvectors associated with the largest eigenvalues of the following stand eigenproblem

$$\left((D-H^T)(D-H^T)^T - (I-W^T)(I-W^T)^T\right)K\alpha = \lambda\alpha \tag{12}$$

Thus, the vector α can also be regarded as the eigenvectors of the matrix $\left((D-H^T)(D-H^T)^T - (I-W^T)(I-W^T)^T\right)K$ associated with the largest eigenvalues. Let $\alpha_1, \alpha_2, \ldots, \alpha_l$ be the solution of equation (12) ordered according to their eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_l$. Thus, for a new document data x , its projection onto U in the feature space F can be calculated as follows

$$x \to y = (U \cdot \varphi(x)) = \sum_{i=1}^{n} \alpha_i k(x_i, x) \tag{13}$$

where y is the low-dimensional feature representation of document data x , and $k(,)$ is the kernel function.

As can be seen, similar to DGPP, our proposed KDGPP also avoids the singularity problem since it does not compute any matrix inverse for dimensionality reduction.

Manifold adaptive kernel function design of KDGPP

From the above statement, we can observe that the kernel function selection is the heart of the proposed KDGPP algorithm since different kernel function will produce different constructions of implicit feature space. The most commonly used kernels include Gaussian kernel and polynomial kernel. However, the nonlinear structure captured by the data-independent kernels may not be consistent with the intrinsic document manifold structure. To improve the classification performance of KDGPP algorithm, we introduce the following manifold adaptive kernel function design method.

Since the data-dependent kernel can be obtained via pairwise constraints[7], we can construct the following similarity matrix T to represent the pairwise constraints

$$T_{ij} = \begin{cases} +1, & (x_i, x_j) \in S \\ -1, & (x_i, x_j) \in D \\ 0, & otherwise \end{cases} \tag{14}$$

where S denotes similar pairwise constraint(the data pairs share the same class), and D denotes dissimilar pairwise constraint(the data pairs have different classes).

In order to model the manifold structure, we construct a nearest neighbour graph G . Each document data point corresponds to a node in G . For each document data point x_i , we find its k nearest neighbors denoted by $N(x_i)$ and put an edge between x_i and its neighbours. Let us define a distance function $h(x)=\|x-x^{(k)}\|$ where $x^{(k)}$ is the kth nearest neighbor of x in G .The weight matrix W associated with graph G is defined as follows:

$$W_{ij}=\begin{cases}\exp\left(-\dfrac{\|x_i-x_j\|^2}{\max\left\{h(x_i)^2,h(x_j)^2\right\}}\right), & \text{if } \|x_i-x_j\|<\max\left\{h(x_i),h(x_j)\right\}\\ 0, & \text{otherwise}\end{cases} \tag{15}$$

Then the normalized graph Laplacian matrix L can be defined as

$$L=I-A^{-1/2}WA^{-1/2} \tag{16}$$

where I is the identity matrix, and A is a diagonal degree matrix given by $A_{ii}=\sum_j W_{ij}$. The normalized graph Laplacian matrix L provides the smoothness penalty on the graph.

Then, by using the graph Laplacian and pairwise constraints, the manifold adaptive non-parametric kernel learning can be formulated the following minimization problem

$$\min_{K\succ 0}\mathrm{Tr}(LK)+C\sum_{(x_i,x_j)\in S\cup D}l\left(T_{ij}K_{ij}\right) \tag{17}$$

where $\mathrm{Tr}(\ )$ is the matrix trace, C is the positive constant to control the tradeoff between the empirical loss $l(\ )$ and the intrinsic data manifold. Since the optimal problem in (17) belongs to typical semi-definite programming (SDP) problem, which can be easily computed using the standard SDP solver SeDuMi. By using the obtained optimal kernel function $k(\ )$, we can greatly improve the classification performance of the proposed KDGPP algorithm.

Efficient computing kernel matrix of KDGPP

In real-world document classification task, the number n of documents is usually large, so performing KDGPP by directly manipulating a $n\times n$ symmetric and positive (semi-) definite kernel matrix is computationally intensive.To reduce the computational demand, in this section, we present an efficient kernel matrix computing algorithm via Nyström method.

The Nyström method is a popular sampling-based low-rank approximation scheme for reducing the computational burdens in handling large kernel matrices. Denote the document data sample set $\mathrm{X}=\{x_i\}_{i=1}^n$ with the with the corresponding $n\times n$ kernel matrix K .Then the Nyström method that randomly chooses a subset $Z=\{z_i\}_{i=1}^m$ of m landmark points will approximate the eigensystem of the full kernel matrix $K\varphi_K=\varphi_K\Lambda_K$ in terms of

$$\varphi_K\approx\sqrt{\frac{m}{n}}E\varphi_Z\Lambda_Z^{-1},\ \Lambda_K\approx\frac{n}{m}\Lambda_Z\,. \tag{18}$$

where $E\in\mathbb{R}^{n\times m}$ with $E_{ij}=k(x_i,z_j)$, and φ_Z , $\Lambda_Z\in\mathbb{R}^{m\times m}$ contain the eigenvectors and eigenvalues of $W_{ij}=k(z_i,z_j)$. By using the approximations in (18), the original kernel matrix K can be reconstructed as

$$K\approx\varphi_K\Lambda_K\varphi_K^T=EW^{-1}E^T\,. \tag{19}$$

According to (19), the the low-rank approximation error $\varepsilon=\left\|K-EW^{-1}E^T\right\|_F$ of the Nyström method depends on the choice of the landmark points z_k , the better the landmark points can encode the data, the lower the resultant low-rank approximation error[8]. Inspired by the above observation and the fact that *k*-means clustering can find a local minimum of the quantization error, we adopt to use the cluster centers obtained from *k*-means as the landmark points in the Nyström low-rank approximation.

In order to illustrate the efficiency of the above kernel matrix computing algorithm, the computational complexity analysis is simply shown in the following. Given a kernel matrix $K\in\mathbb{R}^{n\times n}$, since directly computing K usually take $O(n^3)$ time and $O(n^2)$ memory, while the Nyström method only needs $O(m^2n)$ time and $O(mn)$ memory space where m is much lower than n (i.e., $m\ll n$). Therefore, the proposed Nyström low-rank kernel approximation algorithm can facilitate efficient computation of KDGPP in terms of time and space complexities.

Efficient KDGPP algorithm for document classification

Based on the above statement, we summarize our proposed efficient kernel discriminative geometry preserving projection(KDGPP) algorithm for document classification as follows.

1) Construct the manifold adaptive kernel. Construct a k nearest neighbor G with the weight matrix defined in (15). Calculate the normalized graph Laplacian matrix L as in (16). Calculate the manifold adaptive non-parametric kernel K defined in (17) by using the standard SDP solver SeDuMi.

2) Calculate the DGPP projection matrix. Compute the eigenvectors and eigenvalues for the eigen-problem defined in (12), wherein the kernel matrix K is reconstructed as in (19) via the Nyström method.

3)Obtain the low-dimensional embeddings. Obtain the lower-dimensional feature representations of document data according to (13).

4) Classify in the lower-dimensional feature space. Now, we obtain the lower-dimensional feature representations of the original documents. In the reduced semantic space, we can apply traditional classifier to classify documents into different classes, we adopt the nearest-neighbor classifier for its simplicity in this paper.

Experimental results

In this section, we investigate the performance of our proposed efficient KDGPP algorithm for document classification. The system performance is compared with the LSI algorithm, the NMF algorithm, the LDA algorithm, the LPP[4] algorithm, the MFA algorithm, and the DGPP algorithm, six of the dimensionality reduction algorithms in document classification. We use the same graph structures in the DGPP and KDGPP algorithms, which is built based on the label information. The settings of other algorithms are identical to the description in the corresponding papers.

Two standard document databases were tested in our experiments: Reuters-21578 and 20 Newsgroups (20NG). In all the experiments, the standard TF-IDF weighting scheme is used to generate the feature vector for each document. We simply removed the stop words and no further preprocessing was done. Each document vector is normalized to one and the Euclidean distance is used as the distance measure.

The classification performance is evaluated by comparing the obtained label of each document with that provided by the document corpus. Three metrics, the

classification accuracy, F1-measure, and AUC[9] score are used to measure the classification performance.

In short, the document classification process has three steps. First, we obtain the document subspaces by dimensionality reduction algorithms; then the new documents to be classified are projected into document subspaces; finally, the new documents is classified by the nearest neighbour classifier, where the Euclidean metric is used as our distance measure. Each document database was randomly partitioned into a training set consisting of one half of the whole data set and a testing set consisting of the remainder one half of the whole data set. To reduce the variability, we repeat each experiment ten times on randomly selected training and test datasets and report the average classification results.

Table 1. Classification performance comparison on Reuters-21578

Algorithm	Accuracy	F1	AUC	Times(s)
LSI	78.6%	75.2%	79.1%	23.9
NMF	81.4%	78.5%	82.3%	51.6
LDA	87.5%	84.9%	86.4%	34.8
LPP	88.2%	86.3%	90.3%	42.5
MFA	90.8%	90.2%	93.5%	38.2
DGPP	92.3%	91.5%	94.7%	26.7
KDGPP	95.7%	93.4%	96.2%	23.2

Table 2. Classification performance comparison on 20NG

Algorithm	Accuracy	F1	AUC	Times(s)
LSI	68.5%	62.8%	72.6%	36.3
NMF	76.3%	67.5%	74.5%	67.4
LDA	82.5%	71.4%	78.1%	48.1
LPP	83.7%	72.3%	78.4%	50.6
MFA	89.2%	78.9%	81.7%	43.7
DGPP	90.5%	79.6%	82.5%	39.2
KDGPP	93.6%	81.8%	83.4%	35.1

The experimental results as well as the running time (second) of computing the projection functions for each algorithm on the two databases are reported on the Table 1 and Table 2, respectively. From the experimental results, we can make the following observations.

1)Our proposed KDGPP algorithm consistently outperforms the LSI, LDA,LPP,MFA and DGPP algorithms in terms of classification accuracy, F1-measure, and AUC score. This is because KDGPP considers not only the intraclass geometry but also the discriminative information derived from the interclass samples. In addition, the manifold adaptive kernel function can effectively describe complex nonlinear variations of documents and further enhance the classification performance of KDGPP.

2) The LSI and NMF algorithms perform worse than other algorithms. A possible explanation is that LSI and NMF are unsupervised, they achieve simply object reconstruction and they are not necessarily useful for discrimination and classification tasks.

3) The DGPP and MFA algorithms performs better than LDA and LPP, which demonstrates the importance of utilizing both label information and local manifold structures, as well as characterizing the separability of different classes with the margin criterion.

4) Although DGPP outperforms MFA by considering the discriminative information contained in the nonmarginal samples, it is still a linear technique.Thus DGPP fails to cope with complex nonlinear variations of documents and performs worse than our proposed KDGPP algorithm.

5) Since we adopt an efficient kernel matrix low-rank approximation computing algorithm via Nyström method, which greatly reduce the time and space complexities of computing kernel matrix, our proposed KDGPP algorithm achieves lower running times than other algorithms.

Conclusions

We have proposed a new algorithm for document classification called kernel DGPP(KDGPP). As shown in the experiment results, KDGPP can achieve much better performance than other popular document algorithms.

This work is supported by the National Natural Science Foundation of China under Grant No.70701013.

REFERENCES

[1] Deerwester S.C., Dumais S.T.,Landauer T.K., Furnas G.W., Harshman R.A., Indexing by Latent Semantic Analysis,Journal of the American Society of Information Science,41(1990), No.6, 391-407

[2] Xu W.,Liu X.,Gong Y.,Document Clustering Based on Non-negative Matrix Factorization, Proceedings of SIGIR'03, 2003:267-273

[3] Duda R.O.,Hart P.E.,Stork D.G., Pattern Classification(second edition),Wiley-Interscience, Hoboken, N.J., 2000

[4] Cai D.,He X.,Han J.,Document Clustering Using Locality Preserving Indexing,IEEE Transactions on Knowledge and Data Engineering, 17(2005), No.12,1624-1637

[5] Yan S.,Xu D.,Zhang B.,Zhang H.-J.,Yang Q.,Lin S.,Graph Embedding and Extensions: A General Framework for Dimensionality Reduction,IEEE Transactions on Pattern Analysis and Machine Intelligence,29(2007),No.1,40-51

[6] Song D.,Tao D.,Biologically Inspired Feature Manifold for Scene Classification,IEEE Transactions on Image Processing, 19(2010), No.1,174-184

[7] Zhuang J., Tsang I.W., Hoi S.C.H.,SimpleNPKL: Simple Non-Parametric Kernel Learning, Proceedings of ICML'09,2009:1273-1280

[8] Zhang K.,Kwok J.T.,Clustered Nyström Method for Large Scale Manifold Learning and Dimension Reduction,IEEE Transactions on Neural Networks, 21(2010),No.10,1576-1587

[9] Yang,Y., An Evaluation of Statistical Approaches to Text Categorization,Jounal of Information Retrieval, 1(1999),No.1-2,67-88

***Authors**: Dr. Ziqiang Wang, Lecturer Xia Sun, School of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, china, E-mail: wzqagent@126.com; wzqbox@gmail.com.*