IMMC: Incremental Maximum Margin Criterion

Jun Yan[1] Benyu Zhang[2] Shuicheng Yan[1] Qiang Yang[3] Hua Li[1]
Zheng Chen[2] Wensi Xi[4] Weiguo Fan[4] Wei-Ying Ma[2] Qiansheng Cheng[1]

[1]School of Mathematical Science, Peking University
Beijing, 100871, P. R. China
{yanjun,scyan,lihua}@ math.pku.edu.cn
qcheng@ pku.edu.cn

[2]Microsoft Research Asia
49 Zhichun Road
Beijing, 100080, P. R. China
{byzhang, zhengc, wyma}@microsoft.com

[3]Department of Computer Science
Hong Kong University of Science and Technology
qyang@cs.ust.hk

[4]Virginia Polytechnic Institute and State University
Blacksburg, VA 24060, USA
{xwensi, wfan}@vt.edu

ABSTRACT

Subspace learning approaches have attracted much attention in academia recently. However, the classical batch algorithms no longer satisfy the applications on streaming data or large-scale data. To meet this desirability, Incremental Principal Component Analysis (IPCA) algorithm has been well established, but it is an unsupervised subspace learning approach and is not optimal for general classification tasks, such as face recognition and Web document categorization. In this paper, we propose an incremental supervised subspace learning algorithm, called Incremental Maximum Margin Criterion (IMMC), to infer an adaptive subspace by optimizing the Maximum Margin Criterion. We also present the proof for convergence of the proposed algorithm. Experimental results on both synthetic dataset and real world datasets show that IMMC converges to the similar subspace as that of batch approach.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning
G.1.6 [**Numerical Analysis**]: Constrained Optimization

Keywords

Maximum Margin Criterion (MMC), Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA).

1. INTRODUCTION

In the past decades, machine learning and data mining research has witnessed a growing interest in subspace learning [7] and its applications, such as web document classification and face recognition. Among various subspace learning approaches, linear algorithms are of great interesting due to their efficiency and effectiveness. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are two of the most widely used linear subspace learning algorithms. Furthermore, a novel efficient and robust subspace learning approach namely Maximum Margin Criterion (MMC) [4] was proposed recently. It can outperform PCA and LDA on many classification tasks.

KDD'04, August 22-25, 2004, Seattle, Washington, USA

PCA is an unsupervised subspace learning algorithm. It aims at finding the geometrical structure of data set and projecting the data along the directions with maximal variances. However, it discards the class information, which is significant for classification tasks. On the other hand, LDA is a supervised subspace learning algorithm. It searches for the projection axes on which the data points of different classes are far from each other meanwhile where the data points of the same class are close to each other. Nevertheless, the number of classes limits the available subspace dimension in LDA, and the singularity problem limits the application of LDA. MMC is also a supervised subspace learning algorithm and it has the same goal as LDA. However the computational complexity of MMC is much lower than that of LDA due to the different form of object function.

The original PCA, LDA, and MMC are all batch algorithms, which require that the data must be available in advance and be given once altogether. However, this type of batch algorithms no longer satisfies the applications in which the data are incrementally received from various data sources. Furthermore, when the dimension of the dataset is high, both the computation and storage complexity grow dramatically. Thus, an incremental method is highly desired to compute the adaptive subspace for the data arriving sequentially [5]. Incremental Principal Component Analysis (IPCA) [6] algorithms are designed for such a purpose and have been well established. However, IPCA ignores the valuable class label information. Accordingly, the most representative features derived from IPCA may not be the most discriminant ones. On the other hand, incremental supervised subspace learning algorithms have not been studied sufficiently.

In this paper, we propose an incremental supervised subspace learning algorithm by incrementally optimizing the Maximum Margin Criterion called IMMC. It derives the online adaptive supervised subspace from sequential data samples and incrementally updates the eigenvectors of the criterion matrix. IMMC does not need to reconstruct the criterion matrix when it receives a new sample, thus the computation is very fast. We also prove the convergence of the algorithm in this paper.

The rest of the paper is organized as follows. We introduce some background work on subspace learning, including PCA, IPCA, LDA, and MMC algorithms in Section 2. Then, we present the incremental subspace learning algorithm IMMC and the proof of its convergence in Section 3. Experimental results on the synthetic dataset and the real datasets are shown in Section 4. Finally, we

concluded our work in Section 5, as well as some detailed proof in the appendix.

2. BACKGROUND KNOWLEDGE

Linear subspace learning approaches are widely used in real tasks such as web document classification and face recognition nowadays. It aims at finding a projection matrix, which could efficiently project the data from the original high dimensional feature space to a much lower dimensional representation under a particular criterion. Different criterion will yield different subspace learning algorithm with different properties. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are two most widely used linear subspace learning approaches. Recently, Maximum Margin Criterion, a novel efficient and robust subspace learning approach has also been applied to many real tasks.

2.1 Principal Component Analysis

Suppose that the data sample points $u(1),u(2),...,u(N)$ are d-dimensional vectors, and that U is the sample matrix with $u(i)$ as its i^{th} column. PCA aims to find a subspace whose basis vectors correspond to the directions with maximal variances. It projects the original data into a p-dimensional ($p << d$) subspace. The new low-dimensional feature vector can be computed as $y = W^T u$, where W is the projection matrix and its column vectors correspond to the p leading eigenvectors of the covariance matrix $C = UU^T$. PCA minimizes the reconstruction error in the sense of least square error, and finds out the most representative features. Moreover, PCA is in fact a scalable algorithm since it has effective incremental learning algorithm, which could process large scale streaming data. However, it ignores the class label information; therefore, it is not optimal for general classification tasks.

The computation cost of PCA, which is $O(m^3)$, mainly lies in the SVD processing, where m is the smaller one of the data dimension and the number of samples. Thus, it is difficult or even impossible to conduct PCA on large scale dataset with high dimensional representations.

2.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA), also called Fisher Discriminant Analysis (FDA), was proposed to pursue a low dimensional subspace that can best discriminate the samples from different classes. Suppose $W \in R^{d \times p}$ is the linear projection matrix; LDA aims to maximize the so-called Fisher criterion,

$$J(W) = |W^T S_b W| / |W^T S_w W|,$$

where

$$S_b = \sum_{i=1}^{c} p_i (m_i - m)(m_i - m)^T \quad , \quad S_w = \sum_{i=1}^{c} p_i E(u_i - m_i)(u_i - m_i)^T$$

are called the Inter-class scatter matrix and the Intra-class scatter matrix, respectively, where c is the number of classes, m is the mean of all samples, m_i is the mean of the samples belonging to class i and p_i is the prior probability for a sample belonging to class i. The projection matrix W can be obtained by solving the following generalized eigenvector decomposition problem:

$$S_b w = \lambda S_w w .$$

There are at most c-1 nonzero eigenvalues, so the upper bound of p is c-1; and at least $d + c$ data sample is required to make it possible that S_w is not singular. These constrains limit the application of LDA. Furthermore, it is difficult for LDA to handle large size datasets when the dimension of the feature space is high.

2.3 Incremental PCA

PCA is a batch algorithm. It can not meet the requirement of many real world problems. Incremental learning algorithms have attracted much attention in the past decades. Incremental PCA is a well-studied incremental leaning algorithm. Many types of IPCA have been proposed, and the main difference is the incremental representation of the covariance matrix. The latest version of IPCA [6] with convergence proof is called Candid Covariance-free Incremental Principal Component Analysis (CCIPCA) which does not need to reconstruct the covariance matrix at each iteration of the computation. It was designed based on the assumption that $E\{A(n)\} = A$, where $A \in R^{d \times d}$ is of full rank and positive determined.

2.4 Maximum Margin Criterion

Maximum Margin Criterion (MMC) is a recently proposed feature extraction criterion. This new criterion is general in the sense that combined with a suitable constraint it can actually give rise to the most popular feature extractor in the literature, *i.e.* Linear Discriminant Analysis. Using the same representation as LDA, the goal of MMC is to maximize the criterion $J(W) = W^T (S_b - S_w) W$.

Although both MMC and LDA are supervised subspace learning approaches, the computation of MMC is easier than that of LDA since MMC does not have inverse operation. The projection matrix W can be obtained by solving the following eigenvector decomposition problem:

$$(S_b - S_w) w = \lambda w .$$

When computing, we can notice that the criterion matrix $S_b - S_w$ may even be negative determined.

3. INCREMENTAL MMC

As discussed above, IPCA ignores the class label information. Thus the most representative features found by IPCA may not be the most discriminating ones which make IPCA not being optimal for general classification tasks. It motivates us to design an incremental supervised subspace learning algorithm that can efficiently utilize the label information. In this work, we consider the scenario that maximizes the Maximum Margin Criterion proposed by Li [4] to make the different class centers as far as possible, at the same time make the data points in the same class as close as possible.

In the following subsections, we will introduce the details on how to incrementally maximize the Maximum Margin Criterion. The convergence proof and algorithm summary are also presented.

3.1 Problem Formulation

Denote the projection matrix from original space to the low dimensional space as $W \in R^{d\times p}$. In this work, we propose to incrementally maximize the MMC criterion $J(W) = W^T(S_b - S_w)W$, where S_b and S_w are the inter-class scatter matrix and intra-class scatter matrix respectively. Let C be the covariance matrix. In the above formulation, we exercised freedom to multiply W with some nonzero constant. Thus, we additionally require that W consists of unit vectors, i.e. $W = [w_1, w_2, \cdots w_p]$ and $w_k^T w_k = 1$. Then the optimization problem of the proposed object function $J(W)$ is transformed to the following constrained optimization problem:

$$max \sum_{k=1}^{p} w_k^T (S_b - S_w) w_k \text{ , subject to } w_k^T w_k = 1 \text{ , } k=1,2,\ldots,p \text{ .}$$

Through Lagrangian, it is easy to prove that W is the first k leading eigenvectors of the matrix $S_b - S_w$ and the column vectors of W are orthogonal to each other. It shows that our problem is learning the p leading eigenvector of $S_b - S_w$ incrementally.

3.2 The Leading Eigenvector

Before giving the incremental formulation of MMC, we analyze the criterion matrix $S_b - S_w$ and transform it into a convenient form. Firstly, two lemmas are listed as:

Lemma-1: $S_b + S_w = C$.

Lemma-2: if $\lim_{n\to\infty} a_n = a$ then $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} a_i = a$.

Assume that a data sample sequence is presented as $\{u_{l_n}(n)\}$, where $n=1, 2\ldots$. The goal of MMC is to maximize the Maximum Margin criterion $J(W) = W^T(S_b - S_w)W$, $W \in R^{d\times p}$. Here p is the dimension of the transformed subspace. The Maximum Margin criterion can be transformed as $J(W) = W^T(2S_b - C)W$ from Lemma-1. Then maximizing $J(W)$ means to find the p leading eigenvectors of $2S_b - C$.

The Inter-class scatter matrix of step n after learning from the first n samples can be written as below,

$$S_b(n) = \sum_{j=1}^{c} p_j(n)(m_j(n) - m(n))(m_j(n) - m(n))^T \tag{1}$$

From the fact that $\lim_{n\to\infty} S_b(n) = S_b$ and the lemma-2, we obtain

$$S_b = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} S_b(i) \tag{2}$$

On the other hand,

$$\begin{aligned} C &= E\{(u(n) - m)(u(n) - m)^T\} \\ &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} (u(n) - m(n))(u(n) - m(n))^T \end{aligned} \tag{3}$$

Assume that θ is a positive real number and $I \in R^{d\times d}$ is an identity matrix, if λ is an eigenvalue of matrix A and x is the corresponding eigenvector, then $(A + \theta I)x = Ax + \theta Ix = (\lambda + \theta)x$, i.e. A should have the same eigenvectors with matrix $A + \theta I$. Further more, the order from the largest to the smallest of their corresponding eigenvalues are the same. Therefore, $2S_b - C$ should have the same eigenvectors as $2S_b - C + \theta I$.

From (2) and (3) we have the following conclusion:

$$\begin{aligned} 2S_b - C + \theta I &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} (2S_b(i) - (u(i) - m(i))(u(i) - m(i))^T + \theta I) \\ &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} A(i) = A \end{aligned}$$

where $A(i) = 2S_b(i) - (u(i) - m(i))(u(i) - m(i))^T + \theta I$,

$$A = 2S_b - C + \theta I \text{ .}$$

Notice that we can consider matrix $A(i)$ as a random matrix, in other words we have $E\{A(n)\} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} A(i)$.

The general eigenvector form is $Ax = \lambda x$, where x is the eigenvector of matrix A corresponding to the eigenvalue λ . By replacing matrix A with the Maximum Margin criterion matrix at step n, we obtain an approximate iterative eigenvector computation formulation with $v = \lambda x$:

$$\begin{aligned} v(n) &= \frac{1}{n}\sum_{i=1}^{n} (2S_b(i) - (u(i) - m(i))(u(i) - m(i))^T + \theta I)x(i) \\ &= \frac{1}{n}\sum_{i=1}^{n} (2\sum_{j=1}^{c} p_j(i)\Phi_j(i)\Phi_j(i)^T \\ &\quad -(u(i) - m(i))(u(i) - m(i))^T + \theta I)x(i) \end{aligned} \tag{4}$$

where $\Phi_j(i) = m_j(i) - m(i)$, $v(n)$ is the n^{th} step estimation of v and $x(n)$ is the n^{th} step estimation of x . Once we obtain the estimation of v , eigenvector x can be directly computed as $x = v/\|v\|$. Let $x(i) = v(i-1)/\|v(i-1)\|$, we have the following incremental formulation:

$$\begin{aligned} v(n) &= \frac{n-1}{n} v(n-1) + \frac{1}{n}(2\sum_{j=1}^{c} p_j(n)\Phi_j(n)\Phi_j(n)^T \\ &\quad -(u(n) - m(n))(u(n) - m(n))^T + \theta I)\, v(n-1)/\|v(n-1)\| \\ &= \frac{n-1}{n} v(n-1) + \frac{1}{n}(2\sum_{j=1}^{c} p_j(n)\alpha_j(n)\Phi_j(n) \\ &\quad -\beta(n)(u(n) - m(n)) + \theta v(n-1))\ /\|v(n-1)\| \end{aligned} \tag{5}$$

where $\alpha_j(n) = \Phi_j(n)^T v(n-1)$ and $\beta(n) = (u(n) - m(n))^T v(n-1)$, $j = 1, 2, \ldots, c$. For initialization, we set $v(0)$ be the first data sample.

3.3 Other Eigenvectors

Notice that different eigenvectors are orthogonal to each other. Thus it helps to generate "observations" only in a complementary space for the computation of the higher order eigenvectors. To compute the $(j+1)^{th}$ eigenvector, we first subtract its projection on the estimated j^{th} eigenvector from the data,

$$u_{l_n}^{j+1}(n) = u_{l_n}^{j}(n) - (u_{l_n}^{j}(n)^T v^j(n))v^j(n) \tag{6}$$

where $u_{l_n}^{1}(n) = u_{l_n}(n)$. The same method is used to update $m_i^j(n)$ and $m^j(n)$ $i = 1, 2, \ldots, c$. Since $m_i^j(n)$ and $m^j(n)$ are linear

combinations of $x_{l_i}^j(i)$, where $i=1,2,...,n$, $j=1,2,...,k$, and $l_i \in \{1,2,...,C\}$, Φ_i are linear combination of m_i and m , for convenience, we can only update Φ at each iteration step by:

$$\Phi_{l_n}^{j+1}(n) = \Phi_{l_n}^{j}(n) - (\Phi_{l_n}^{j}(n)^T v^j(n))v^j(n) \quad (7)$$

In this way, the time-consuming orthonormalization is avoided and the orthogonal is always enforced when the convergence is reached, although not exactly so at early stages.

Through the projection procedure using (6) (7) at each step, we can get the eigenvectors of Maximum Margin criterion matrix one by one. It is much more efficient in comparison to the time-consuming orthonormalization process.

3.4 Convergence Proof Summary

The full algorithm consists of updating (5), (6) and (7) at each iteration. Theorem-1 shown as below guarantees the convergence of the proposed Incremental Maximum Margin Criterion algorithm when the selected positive real number θ makes the matrix $2S_b - C + \theta I$ is non-negative determined.

A similar theorem with proof can be found in [8]. Our convergence proof for eigenvectors except the largest one is the same as it. We just give out the proof summary and ignore the parts which are the same as in [8].

Theorem-1: If matrices sequence $\{A(n)\}$, $\|A(n)\| < \infty$ converge to a matrix $A \in R^{d\times d}$, i.e. $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} A(i) = A$, where A is nonnegative determined matrix and $\|A\| < \infty$, the eigenvalues of A satisfy $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d \geq 0$, then the iterative process converges to the maximum eigenvalue of matrix A multiplied by the corresponding eigenvector.

$$v(n) = \frac{n-1}{n} v(n-1) + \frac{1}{n} A(n) \frac{v(n-1)}{\|v(n-1)\|} \quad (8)$$

Theorem-2: Suppose $v(n) = v(n-1) + a_n h(v(n-1)) + a_n b_n + a_n \xi_n$. If A1 to A4 are all satisfied, let $\{v(n)\}$ be bounded w.p.1.

A 1 $h(\cdot)$ is a continuous R^d valued function on R^d .

A 2 $\{b_n\}$ is a bounded sequence of R^d valued random variables such that $b_n \to 0$ when $n \to \infty$.

A 3 $\{a_n\}$ is a sequence of positive real numbers such that $b_n \to 0$, $\sum_n a_n = \infty$.

A 4 $\{\xi_n\}$ is a sequence of R^d valued random variables and such that for some $T > 0$ and each $\varepsilon > 0$

$$\lim_{n\to\infty} p\{\sup_{j\geq n} \max_{t\leq T} \left|\sum_{i=m(jT)}^{m(jT+t-1)} a_i \xi_i\right| \geq \varepsilon\} = 0 .$$

Let v^* be a locally asymptotically stable (in the sense of Liapunov) solution to equation $dX/dt = h(X)$ with the domain of attraction $DA(v^*)$ and there is a compact set $H \in DA(v^*)$ such that $v(n) \in H$ infinitely often, we have $v(n) \to v^*$ as $n \to \infty$. (The origin form of this theorem and its proof can be found in [2].)

Theorem-3: Let $v(t) \to v^*$ be a locally asymptotically stable (in the sense of Liapunov) solution to the Ordinary Differential Equation as bellow:

$$\frac{dv}{dt} = \left(\frac{A}{\|v\|} - I\right)v \quad (9)$$

where $A \in R^{d\times d}$ is a nonnegative determined matrix, $v \in R^d$ and the eigenvectors of A satisfies $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. Then $v(t)$ converges to $\lambda_1 e_1$, e_1 is the eigenvector corresponding to λ_1 .

The combination of theorem-2 and theorem-3 gives out the proof of theorem-1 which is in fact the convergence proof of our proposed Incremental Maximum Margin Criterion algorithm. It is easy to prove that the proposed algorithm satisfies the conditions of theorem 2. A 1, A 2 and A 3 are naturally satisfied and A.4 is satisfied due to lemma-3. The combination of Lemma-4 and the proof of [8] ends the proof of theorem-3, *i.e.* the convergence proof of our proposed algorithm.

Lemma-3: $v(n)$ is bounded, if $v(0)$ is bounded.

The proof of lemma-3 could be found in the appendix.

Lemma-4: $A \in R^{d\times d}$ is a nonnegative determined matrix, $rank(A) = m$ and $m \leq d$, $\{e_i\}$ $i = 1,2,\cdots m$ are eigenvectors corresponding to non-zero eigenvalues of *A,* if we expend $\{e_i\}$ to a normalized orthogonal basis of R^d , $e_1, e_2, \cdots e_m, e_{m+1}, \cdots e_d$, then $Ae_j = 0$, $j = m+1, \cdots, d$.

Proof: set $y_j = Ae_j \neq 0$, $j = m+1, \cdots, d$, we have $y_j \in span\{e_i; i = 1,2,\cdots,m\}$ and it conflicts with the fact that $e_j \perp span\{e_i; i = 1,2,\cdots,m\}$, this ends the proof of lemma-4.

The time complexity of IMMC to train N input samples is $O(Ncdp)$, where c is the number of classes, d is the dimension of the original data space and p is the target dimension, which is linear with each factor. Furthermore, when handling each input sample, IMMC only need to keep the learned eigen-space and several first order statistics of the past samples, such as the mean and the counts. Hence, IMMC is able to handle large scale and continuous data stream.

4. EXPERIMENTAL RESULTS

We performed three sets of experiments. Firstly, we used synthetic data to illustrate the subspaces learned by IMMC, LDA, and PCA intuitively. Secondly, we applied IMMC on some UCI subsets [1], and compared the results with the batch MMC approach that conducted by SVD, whose time complexity is $O(m^3)$, where m is the smaller number of the data dimension and the number of samples. Since the classification performance of MMC such as LDA has been discussed when it was proposed, we only focus on the convergence performance of IMMC to the batch MMC algorithm on UCI dataset. In the third dataset, the Reuters Corpus Volume 1 (RCV1) [3], a large scale dataset whose dimension is about 300,000, was used. We measured the performance of our algorithm by F1 value on it because the dataset is too large to conduct the batch MMC on it.

4.1 Synthetic dataset

We generated a 2-dimension dataset of 2 classes. Each class consists of 50 samples from normal distribution with means (0, 1) and (0,-2), respectively; and the covariance matrices are *diag*(1, 25) and *diag*(2, 25). Figure 1 shows a scatter plot of the data set. The two straight lines are subspaces found by IMMC and PCA. Since the subspace found by MMC is the same as subspace by LDA in the case, we did not give out the LDA subspace.

Since $\|v - v'\| = 2(1 - v \cdot v')$, and $v = v'$ iff $v \cdot v' = 1$, the correlation between two unit eigenvectors is represented by their inner product, and the larger the inner product is, the more similar the two eigenvectors are. Let us analyze this dataset to show the convergence ability of IMMC. For this toy data the eigenvalues of $A = 2S_b - C + \theta I$ are *0.25* and *-84.42* and the corresponding eigenvectors are (0,-1) and (-1, 0). We choose $\theta = 85$ to make sure that the criterion matrix is nonnegative determined. Figure 2 shows the convergence curve of our algorithm through inner product.

4.2 Real World Data

UCI machine learning dataset is a repository of databases, domain theories and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. Balance Scale Data was generated to model psychological experimental results.

The number of instances is 625 and the number of attributes is 4. There are three classes (49 balanced, 288 left, 288 right.). For this Balance Scale data set, the eigenvalues of $A = 2S_b - C + \theta I$ are *-2.0*, *-2.0*, *-1.9774*, and *0.7067*. We choose $\theta = 2.0$ to make sure the criterion matrix is nonnegative determined. Figure 3 shows the inner products of directions found by IMMC and CCIPCA [6].

In order to demonstrate the performance of IMMC on relative large scale data, we choose the Ionosphere database (figure 4). This radar data was collected by a system in Goose Bay, Labrador. This system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts.

The number of instances is 351, and the number of attributes is 34 plus the class attribute. All 34 predictor attributes are continuous and the 35th attribute is either "good" or "bad" according to the definition. Since the smallest eigenvalue of this data set is very close to zero, we try taking the parameter $\theta = 0$ in this experiment.

Unfortunately, some experimental results show that IMMC could not be used on some special data set, if the criterion matrix $A = 2S_b - C$ is negative determined. This difficulty motivates us to propose a weighted Maximum Margin Criterion $A = S_b - \varepsilon S_w$. Some advanced experiments show that the classical MMC ($\varepsilon = 1$) is not usually optimal for classification tasks. In other words, a proper ε could improve the performance of MMC and it could make sure that the criterion matrix is nonnegative determined. Then we could make the criterion matrix nonnegative determined by giving a smaller ε instead of parameter θ. To demonstrate the performance of IWMMC on a large scale dataset, we tested our algorithm on the Reuters Corpus Volume 1 (RCV1).

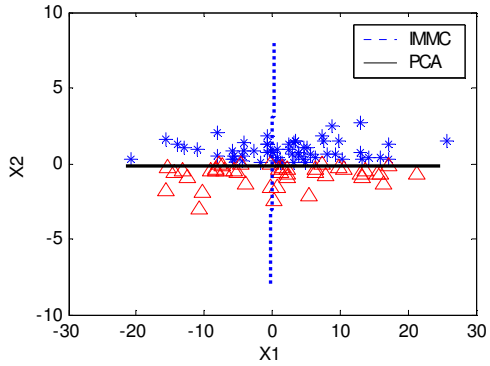

Figure 1 Subspace learned by IMMC and PCA

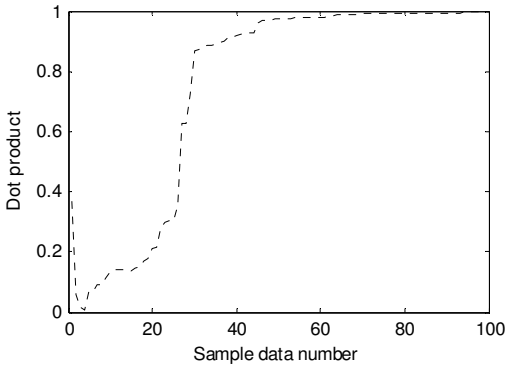

Figure 2 Correlation between eigen-space of IMMC and batch MMC on synthetic dataset

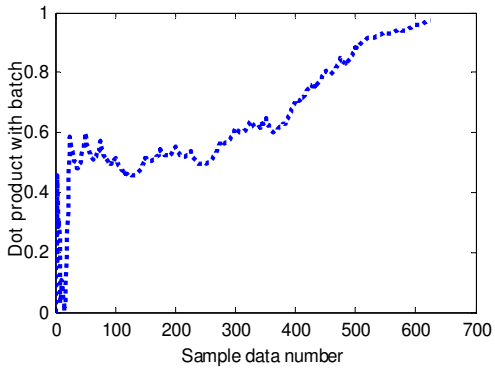

Figure 3 Inner product of first eigenvector with batch approaches by IMMC for BS

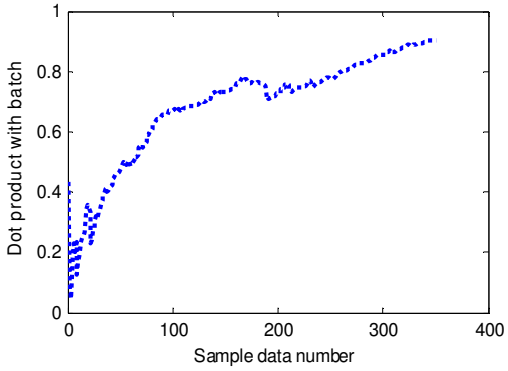

Figure 4 Inner product of first eigenvector with batch approaches by IMMC for Ionosphere $\theta = 0$

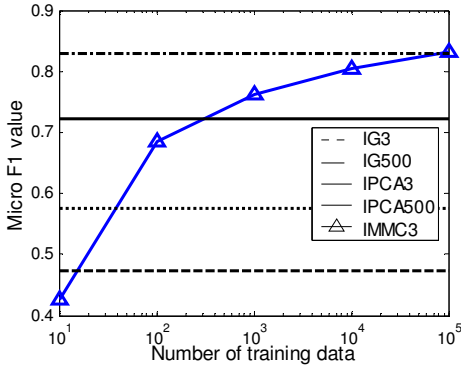

Figure 5 F1 value of incremental weighted MMC

The dimension of each data sample is about 300,000. We chose the data samples with the highest four topic codes in the "Topic Codes" hierarchy, which contains 789,670 documents. Then we applied a five-fold cross validation on the data. We split them into five equal-sized subsets and in each experiment four of them are used as the training set and the remaining one is left as the test set. Figure 5 shows the F1 value of different subspace learning approach by SVM classifier, where the number denotes the subspace dimension. For example, IG3 represents the 3-dimensional subspace calculated by Information Gain. It shows that IWMMC ($\varepsilon = \theta = 0$) outperforms Information Gain and IPCA which could also be conducted on large scale dataset.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed an incremental supervised subspace learning algorithm, called Incremental Maximum Margin Criterion (IMMC), which is a challenging issue of computing dominating eigenvectors and eigenvalues from incrementally arriving stream without storing the knowing data in advance. The proposed IMMC algorithm is effective and has fast convergence rate and low computational complexity. It can be theoretically proved that IMMC can find out the same subspace as batch MMC does. Moreover, batch MMC can approach LDA when there is a suitable constraint. But it remains unsolved that how to estimate and choose the parameter to make sure the criterion matrix is nonnegative determined. In the future work, we intend to give a rational function of θ to make IMMC more stable.

6. ACKNOWLEDGMENTS

This work is accomplished in Microsoft Research Asia. The authors thank Ning Liu (the graduate student of Tsinghua University) for helpful discussions. Q. Yang thanks Hong Kong RGC for their support.

7. REFERENCES

[1] C.L., B. and C.J., M. UCI Repository of machine learning databases Irvine. *CA: University of California, Department of Information and Computer Science.*

[2] Kushner, H.J. and Clark, D.S. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, New York, 1978.

[3] Lewis, D., Yang, Y., Rose, T. and Li, F. *RCV1: A new benchmark collection for text categorization research.* Journal of Machine Learning Research.

[4] Li, H., Jiang, T. and Zhang, K., Efficient and Robust Feature Extraction by Maximum Margin Criterion. In *Proceedings of the Advances in Neural Information Processing Systems 16*, (Vancouver, Canada, 2004), MIT Press.

[5] Liu, R.-L. and Lu, Y.-L., Incremental Context Mining for Adaptive Document Classification. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (Edmonton, Alberta, Canada, 2002), 599-604.

[6] Weng, J., Zhang, Y. and Hwang, W.-S. *Candid Covariance-free Incremental Principal Component Analysis.* IEEE Trans .Pattern Analysis and Machine Intelligence, 25 (8). 1034-1040.

[7] Yu, L. and Liu, H., Efficiently Handling Feature Redundancy in High-Dimensional Data. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (Washington DC., 2003), 685-690.

[8] Zhang, Y. and Weng, J. *Convergence analysis of complementary candid incremental principal component Analysis*, Michigan State University, 2001.

8. APPENDIX

Lemma-3: $v(n)$ is bounded, if $v(0)$ is bounded.

Proof:

$$\begin{aligned}\|v(n)\| &= \left\| \frac{n-1}{n} v(n-1) + \frac{1}{n} A(n) \frac{v(n-1)}{\|v(n-1)\|} \right\| \\ &\le \frac{n-1}{n} \|v(n-1)\| + \frac{1}{n} \left\| A(n) - \frac{1}{n-1} \sum_{i=1}^{n-1} A(i) \right\| + \frac{1}{n} \left\| \frac{1}{n-1} \sum_{i=1}^{n-1} A(i) \right\|\end{aligned}$$

Since $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} A(i) = A$, from Cauchy Convergence Theorem, $\forall \varepsilon > 0, \exists N_1$, s.t.

$$\left\| \frac{1}{n} \sum_{i=1}^{n} A(i) - \frac{1}{n-1} \sum_{i=1}^{n-1} A(i) \right\| < \frac{\varepsilon}{2}, \text{ when } n \ge N_1 .$$

From the fact that $\|A\| < \infty$ and $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} A(i) = A$, we know that, $\left\| \frac{1}{n} \sum_{i=1}^{n} A(i) \right\|$ is bounded, there must $\exists N_2$ s.t. $\frac{1}{n} \left\| \frac{1}{n-1} \sum_{i=1}^{n-1} A(i) \right\| < \frac{\varepsilon}{2}$ when $n \ge N_2$.

Let $N = \max\{N_1, N_2\}$, then $\|v(n)\| \le \|v(n-1)\| + \varepsilon$ when $n \ge N$. Since we can choose ε freely, we can draw the conclusion that $\|v(n)\| \le \|v(n-1)\|$. when $n \ge N$.

Since $\|v(0)\| < \infty$, When $n \le N$

$$\begin{aligned}\|v(n)\| &\le \frac{n-1}{n} \|v(n-1)\| + \frac{1}{n} \|A(n)\| \le \cdots \\ &\le \|v(0)\| + \frac{1}{n} (\|A(n)\| + \|A(n-1)\| + \cdots \|A(1)\|) < \infty\end{aligned}$$

So $\|v(n)\|$ is bounded when $n \le N$ and $\|v(n)\| \le \|v(n-1)\|$ when $n \ge N$, i.e. $\|v(n)\|$ is bounded, $n = 1, 2, \cdots$. Notice this implies that $\|v(n)\|$ w.p.1.

End of proof.