QR-분해를 이용한 효율적인 차원 감소 방법과 문서 분류에의 응용

이문휘, 박정희
컴퓨터공학과, 충남대학교
{moone81, cheonghee}@cnu.ac.kr

Efficient dimension reduction using QR-decomposition and its application to text categorization

Moonhwi Lee, Cheong Hee Park
Dept. of Computer Science and Engineering, Chungnam National University

요 약

LDA는 그룹간 간격을 최대화하고 그룹내 분산을 최소화하는 선형변환을 구함으로써 차원 감소된 공간에서 분별력(classification performance)을 높이는 선형 차원 감소 방법이다. 본 논문에서는 저샘플 문제(undersampled problem)에서 LDA를 적용할 수 있도록 QR-분해를 이용한 효율적인 차원 감소 방법을 제안한다. 특히 제안되는 방법은 문서 분류 문제에서처럼 한 문서가 몇 개의 카테고리에 중복적으로 속하는 경우 등 데이터의 독립성이 보장되지 않는 경우에도 효과적으로 적용될 수 있다는 장점이 있다.

1. 서론

차원 감소란 고차원 데이터에서 최적의 특징들을 추출함으로써 데이터 공간의 차원을 줄이는 작업이다. 이는 변형된 저차원 공간에서 주요 작업들을 빠르게 하고 잡음의 영향을 줄일 수 있는 효과적인 데이터 처리를 위한 중요한 선행 과정이 되고 있다. 주성분분석법(Principal Component Analysis, PCA)과 선형판별분석법(Linear Discriminant Analysis, LDA)은 널리 사용되는 전통적인 차원 감소 방법들이다. PCA는 데이터의 분산이 가장 큰 방향으로 정사영 함으로써 차원 감소로 인한 정보 손실을 최소화 하는 반면, LDA는 그룹간 간격을 최대화하고 그룹내 분산을 최소화하는 선형변환을 구함으로써 차원 감소된 공간에서 분별력(classification performance)을 높이는 것을 목적으로 한다[1]. 그러나 데이터의 수가 데이터 차원보다 적은 저샘플 문제(undersampled problem)에서는 산포행렬(scatter matrix)의 특이성(singularity) 때문에 LDA를 적용하기 어렵다[2]. 최근에 이러한 저샘플 문제에 적용 가능한 일반화된 LDA 알고리즘들이 제안되어졌다. 그러나 대부분의 일반화된 LDA 방법들에서 사용되는 SVD(singular value decomposition)는 계산 복잡도가 높고 메모리 요구가 크다.

최근 계산의 효율성을 위해 SVD 대신 Gram-Schmidt 직교화에 의한 QR-분해를 이용하여 계산복잡도를 낮추는 LDA 알고리즘인 GSLDA가 제안되었다[3]. 그러나 GSLDA는 데이터 항목들의 독립성을 전제로 하기 때문에 데이터들이 독립적이지 않거나 문서 분류 문제에서처럼 한 문서가 몇 개의 카테고리에 중복적으로 속하는 경우 등 데이터의 독립성이 보장되지 않는 실제데이터에 적용할 수 없다는 단점이 있다.

본 논문에서는 데이터의 독립성에 관계 없이 저샘플 문제에 적용할 수 있는, QR-분해를 이용한 효과적인 차원 감소 방법을 제안하고 문서 분류 문제에 적용하여 제안된 알고리즘의 효용성을 입증한다. 2절에서는 LDA에 의한 차원감소 방법과 Gram-Schmidt 직교화를 이용한 GSLDA에 대해 간단히 설명하고 3절에서 QR-분해를 이용한 차원 감소 방법을 제안한다. 마지막으로 4절에서는 제안 되어진 방법과 기존의 방법들에 대한 비교 실험을 통해 성능을 입증한다.

2. 선형 판별 분석(Linear Discriminant Analysis)

r개의 그룹으로 분류되어진 원소들을 가지는 데이터 집합 A를 다음과 같이 나타내자.

$A = \{a_1, \Lambda, a_n\} = \{a_i^j : 1 \le i \le n_j,\ 1 \le j \le r\}$

각 그룹 j (1 <= j <= r)는 n_j 개의 원소들을 가지며 데이터의 총 개수는 $n = \sum_{1 \le j \le r} n_j$ 이다. 그룹간 또는 그룹내 데이터 분산을 나타내기 위해 그룹간 분산 행렬(between-class scatter matrix) S_b, 그룹내 분산 행렬(within-class scatter matrix) S_w 그리고 총 분산 행렬(total scatter matrix) S_t 가 아래와 같이 정의된다.

$$S_b = \sum_{1 \le j \le r} n_j (c_j - c)(c_j - c)^T,$$

$$S_w = \sum_{1 \le j \le r} \sum_{1 \le i \le n_j} (a_i^j - c_j)(a_i^j - c_j)^T, \qquad (1)$$

$$S_t = \sum_{1 \le j \le r} \sum_{1 \le i \le n_j} (a_i^j - c)(a_i^j - c)^T.$$

위에서, $c_j = \sum_{1\le i\le n_j} a_i^j \Big/ n_j$ 와 $c = \sum_{1\le j\le r}\sum_{1\le i\le n_j} a_i^j \Big/ n$ 는 각 그룹의 중심과 전체 중심을 나타낸다. LDA에서 그룹간 간격을 최대화 하고 그룹내 분산을 최소화 하는 선형변환을 고유값 문제 $S_w^{-1}S_b x = \lambda x$ 를 해결함으로써 구할 수 있음이 알려져 있다[1]. 그러나 데이터의 개수가 데이터의 차원보다 적다면, 분산행렬이 특이 행렬이 되어 전통적인 LDA를 적용할 수 없게 된다.

Zheng 등은, 저샘플 문제 해결을 위한 최적의 선형변환을 위한 벡터들이 $range(S_t)\cap null(S_w)$ 로부터 선택될 수 있다고 주장했다[3]. 그들은 modified Gram-Schmidt 직교화(MGS)를 사용하는 계산적으로 효과적인 방법인 GSLDA를 제안했다. 데이터 항목들이 독립적인 경우에 $range(S_t)\cap null(S_w)$ 은 r-1개, 즉, 그룹들의 수-1개의 직교 벡터들에 의한 기저를 가지게 되며, 이러한 직교 기저에 의한 선형변환은 r-1 차원으로 감소된 공간에서 $null(S_w)$ 성질에 의해 데이터들의 그룹내 분산을 극소화 한다. GSLDA의 자세한 알고리즘을 위해 논문 [3]을 참조하라.

그러나 데이터들이 독립적이지 않거나 두개 이상의 그룹들에 속하는 데이터 항목들이 존재할 경우 데이터 독립성을 전제로 하는 GSLDA는 분류 성능이 크게 떨어지거나 적용이 불가능하게 된다. 다음 절에서 QR-분해를 이용하여 계산 복잡도를 줄이면서도 데이터의 독립성을 필요로 하지 않는 차원 감소 방법을 제안한다.

3. QR-분해를 이용한 차원 감소 방법

(1)에서 정의된 분산 행렬들은 아래의 H_w, H_t, H_b 를 이용해 $S_w = H_w H_w^T$, $S_t = H_t H_t^T$, $S_b = H_b H_b^T$ 와 같이 계산 될 수 있다.

$$H_w = [a_1^1 - c_1, \Lambda, a_{n1}^1 - c_1, \Lambda, a_1^r - c_r, \Lambda, a_{nr}^r - c_r],$$
$$H_t = [a_1^1 - c, \Lambda, a_{n1}^1 - c, \Lambda, a_1^r - c, \Lambda, a_{nr}^r - c], \quad (2)$$
$$H_b = [\sqrt{n_1}(c_1 - c), \Lambda, \sqrt{n_r}(c_r - c)]$$

$Range(S_w)$ 와 $null(S_w)$ 는 직교보공간(orthogonal complement)으로서 모든 벡터는 $range(S_w)$ 와 $null(S_w)$ 에 속하는 성분들의 합으로 유일하게 표현될 수 있다[4]. 먼저 $range(S_w)$ 의 직교 기저를 구하기 위해 H_w 에 열치환(column pivoting)을 가진 QR-분해를 적용한다.

$$H_w \Pi_r = Q_w R_w \quad (3)$$

여기서 Π_r 은 열치환 행렬이고 Q_w 는 $range(S_w)$ 의 직교 기저가 된다. H_t 를 $null(S_w)$ 로의 정사영과 $range(S_w)$ 로의 정사영의 direct sum으로 표현할 수 있다. 즉, K 를 H_t 의 $null(S_w)$ 로의 정사영이라 할 때,

$H_t = K \oplus Q_w Q_w^T H_t$ 이다.

따라서, $K \subseteq range(S_t)\cap null(S_w)$ 이며

$K = H_t - Q_w Q_w^T H_t$ 에 의해 구할 수 있다.

다시 한번 열치환을 가진 QR-분해를 K 에 적용함으로써 K 의 직교 기저 Q_t 를 얻는다.

이제 $Q_t^T H_b$ 의 SVD, 즉 $Q_t^T H_b = U_b \Sigma_b V_b^T$ 에 의해 $range(S_t)\cap null(S_w)$ 에서 그룹간 분산을 최대화 하는 변환 행렬 W 를 U_b 의 처음 r-1개의 열에서 얻고, $W^T Q_t^T$ 는 LDA의 해가 된다.

데이터들의 독립성이 약해질수록 $range(S_t)$ 의 rank는 작아진다. 따라서 $range(S_t)\cap null(S_w)$ 의 rank 또한 심하게 작아지는 현상이 발생할 수 있으며 이로 인해 $null(S_w)$ 에서 해를 구하는 일반화된 LDA 알고리즘들은 성능이 크게 떨어지게 된다. 위에서 제안된 방법도 $null(S_w)$ 에서 해를 구하기 때문에 같은 문제를 가진다.

이를 극복하기 위해 우리는 데이터의 변형(perturbation)을 이용 하고자 한다.

(2)에서 $H_t = H_w + \tilde{H}_b$,

$\tilde{H}_b = [c_1 - c, \Lambda, c_1 - c, \Lambda\ c_r - c, \Lambda\ c_r - c)]$

가 성립함을 쉽게 확인할 수 있다.

식 (3)에서 얻어진 치환 행렬 Π_r 을 이용하여 H_t 를 $H_t \approx H_w \Pi_r + \tilde{H}_b$ 로 변형하고 이를 이용해 다음과 같이 K 를 구한다.

$$K = H_t - Q_w Q_w^T H_t \quad (4)$$
$$\approx H_t - Q_w Q_w^T (H_w \Pi_r + \tilde{H}_b)$$

이 방법은 또한 $null(S_w)$ 에서 해를 구함으로써 오는 과적합(overfitting)을 줄이는 효과를 얻을 수 있다.

제안된 방법은 LDA에서 일반적으로 사용되는 SVD에 비해 적은 연산을 요구하는 QR-분해를 이용함으로써 시간 복잡도와 메모리 요구를 줄일 수 있다. 또한 GSLDA에서 요구되는 전제조건인 데이터 독립성을 필요로 하지 않는다는 장점을 가진다. 다음 절에서는 문서 분류 문제에 제안된 알고리즘을 적용하여 다른 일반화된 LDA 알고리즘들과 성능을 비교한다.

4. 실험 결과

문서 분류 문제는 미리 몇 개의 그룹으로 분류된 문서 정보를 바탕으로 새 문서의 그룹 라벨을 할당하는 작업이다. 문서들은 term-document matrix로 표현되며, 각 문서는 열벡터로 표현되고 행벡터의 성분은 문서에 나타난 단어의 빈도를 나타낸다[5]. term-document matrix에 의한 문서 데이터의 고차원 표현은 차원 감소 과정의 필요성을 부여한다.

모든 실험은 3.2Ghz CPU, 2G RAM의 Linux 시스템에서 Matlab7을 이용하여 실행하였다.

첫번째 실험을 위해, web[6]에 공개 되어 있는 7종류의 문서 데이터 셋을 사용하여 제안된 알고리즘을 DirectLDA[7], GSLDA[3], twostageLDA[8]와 비교하였다. 데이터 셋에 관한 상세 정보는 표 1에 나타내었다.

데이터 셋은 실험 가능한 크기로 줄이기 위해 빈도수 5 이하인 용어를 제거하는 등의 전처리 과정을 거쳤다. 각 데이터 셋에 대해 학습 데이터와 테스트 데이터로 1:1 비율로 랜덤하게 나누어 분류 정확도를 구했으며 이를 10회 반복한 평균 정확도를 표 2에 나타내었다. 먼저 학습 데이터를 이용해 차원 감소 선형변환을 구하고, 이에 의해 학습 데이터와 테스트 데이터를 저차원 공간으로 보낸 후 1-nearest classifier를 사용하여 분류 정확도를 측정하였다.

표 1. 데이터 셋 정보

Data	re0	re1	wap	hitech	k1a	k1b	la1
Dim	2886	3758	8460	13170	13879	13879	17273
No. Data	1504	1657	1650	2301	2340	2340	3204
Classes	13	25	20	6	20	6	6

표 2. 분류 결과(실험 1)

	dlda[7]	gslda[3]	twostage lda[8]	제안된 방법
re0	0.799	0.665	0.724	**0.816**
re1	0.838	0.756	0.821	**0.842**
wap	0.758	0.780	**0.781**	0.778
hitech	0.704	**0.705**	**0.705**	**0.705**
k1a	0.796	0.799	**0.816**	0.803
k1b	0.959	0.959	**0.962**	0.955
la1	0.844	0.858	**0.862**	0.846

두번째 실험에서는 중복 라벨을 갖는 reuter 데이터 셋을 사용하였다[5]. 이 실험에서 총 90개의 그룹들 중 가장 큰 크기를 가지는 10개의 그룹들을 사용하였다. 문서의 중복도를 달리하기 위해 다음과 같이 데이터 셋을 구성하였다. 학습 데이터 셋에 있는 데이터들 중 약 $\frac{1}{10}$이 두 개 이상의 그룹에 속하는 문서들이다. 이러한 중복되는 라벨을 가지는 문서들(B)과 그렇지 않은 문서들(A)을 분리한 후, 적절한 비율로 변화를 주어 A와 B로부터 랜덤하게 문서를 선택하여 총 9개의 데이터 셋을 구성하였다. 표 3은 데이터 구성 정보와 비교된 알고리즘들의 분류 정확도를 보여준다.

표 3. 분류 결과(실험 2)

A	B	dlda[7]	gslda[3]	twostage lda[8]	제안된 방법
$A * \frac{1}{4}$	B * 0	0.858	0.788	0.814	**0.879**
	B * 1/2	0.859	0.638	0.790	**0.878**
	B * 1	0.864	0.615	0.766	**0.881**
$A * \frac{1}{2}$	0	0.876	0.391	0.556	**0.898**
	1/2	0.883	0.565	0.562	**0.898**
	1	0.881	0.525	0.316	**0.900**
A * 1	0	0.885	0.381	-	**0.899**
	1/2	0.881	0.442	-	**0.896**
	1	0.894	0.390	-	**0.904**

- 로 표시된 부분은 메모리 부족에 의한 실험 불가능을 나타낸다.

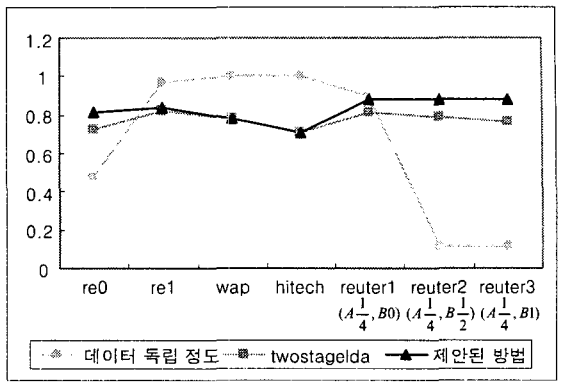

그림 1. twostagelda와 제안된 방법에 의한 분류 정확도와 데이터 독립성의 관계

그림 1은 각 데이터에 대한 twostagelda와 제안된 방법의 분류 정확도를 데이터의 독립성과 비교하여 보여준다. 데이터의 독립성은

$rank(H_b$의 $range(S_t) \cap null(S_w)$로의 정사영$) / r - 1$로 측정되었으며 1에 가까울수록 데이터 샘플들이 독립적임을 의미한다.

위의 실험결과들은 제안된 방법이 독립적인 데이터에 대해서 경쟁력 있는 분류 성능을 가지며 re0, re1과 같이 데이터 독립성이 보장되지 않거나 reuter 데이터처럼 중복 라벨을 갖는 경우에는 월등한 분류 성능을 나타냄을 입증한다.

5. 참고 문헌

[1] R.O. Duda, P.E. Hart, D.G. Stork, "Pattern classification, " Wiley-interscience, New York, 2001.

[2] L. Chen, H. Liao, M. Ko, J. Lin, G. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," Pattern Recognition 33 (10), pp1713-1726, 2000.

[3] W. Zheng, C. Zou, L. Zhao, "Real-Time Face Recognition Using Gram-Schmidt Orthogonalization for LDA," ICPR'04, pp.403-406, 2004.

[4] L.E. Mansfield, "Linear algebra with geometric applications," Dekker, 1976.

[5] H. Kim, P. Howland, H. Park, "Text classification using support vector machines with dimension reduction," Proceedings of Text Mining Workshop of SDM03, San Francisco, CA, May 1-3, 2003.

[6] http://glaros.dtc.umn.edu/gkhome/views/cluto

[7] H Yu, J. Yang, "A direct LDA algorithm for high-dimensional data - with application to face recognition," Pattern Recognition 34, pp2067-2070, 2001.

[8] J. Yang, J.-Y. Yang, "Why can LDA be performed in PCA transformed space?," Pattern Recognition 36, pp685-690, 2003.