http://dx.doi.org/10.3745/KIPSTD.2012.19D.2.147

앙상블 접근법을 이용한 반감독 차원 감소 방법

박 정 희†

요 약

클래스들 간의 거리를 최대화시키는 사영 방향을 구하는 감독차원감소 방법인 선형판별분석법(LDA)은 클래스 정보를 가진 데이터의 수가 매우 적을 때 성능이 급격히 저하되는 경향이 있다. 이러한 경우 상대적으로 저렴한 비용으로 얻을 수 있는 클래스 라벨 정보가 없는 데이터를 활용할 수 있는 반감독 차원 감소법이 사용될 수 있다. 그러나 통계적 차원 감소법에서 흔히 사용되는 행렬연산은 많은 양의 데이터를 사용하는데 메모리와 처리시간에서 한계가 있고, 적은 수의 라벨드 데이터(labeled data)에 비해 너무나 많은 언라벨드 데이터(unlabeled data)의 사용은 처리 시간의 증가에 비해 오히려 성능감소를 가져올 수 있다. 이러한 문제들을 극복하기 위해 앙상블 접근법을 이용한 반감독 차원 감소 방법을 제안한다. 문서분류 문제에서의 실험결과를 통해 제안한 방법의 성능을 입증한다.

키워드 : 선형판별분석, 반감독차원감소, 문서분류, 앙상블 방법

A Semi-supervised Dimension Reduction Method Using Ensemble Approach

Park, Cheong Hee†

ABSTRACT

While LDA is a supervised dimension reduction method which finds projective directions to maximize separability between classes, the performance of LDA is severely degraded when the number of labeled data is small. Recently semi-supervised dimension reduction methods have been proposed which utilize abundant unlabeled data and overcome the shortage of labeled data. However, matrix computation usually used in statistical dimension reduction methods becomes hindrance to make the utilization of a large number of unlabeled data difficult, and moreover too much information from unlabeled data may not so helpful compared to the increase of its processing time. In order to solve these problems, we propose an ensemble approach for semi-supervised dimension reduction. Extensive experimental results in text classification demonstrates the effectiveness of the proposed method.

Keywords : Linear Discriminant Analysis, Semi-Supervised Dimension Reduction, Text Classification, Ensemble Method

1. 서 론

차원 감소는 분류나 군집 알고리즘의 성능을 향상시킬 수 있는 효과적인 전처리 과정으로 데이터의 클래스 정보 활용 여부에 따라 감독학습(supervised learning)과 무감독학습(unsupervised learning)으로 구분된다. 주성분분석법(Principal component analysis, PCA)는 분산을 최대화시키는 방향으로 데이터들을 사영하는 대표적인 무감독 차원 감소 방법인데 반해, 선형판별분석법(Linear discriminant analysis, LDA)는 클래스 간 거리를 최대화하면서 동시에 클래스 내 분산을 최소화하는 방향을 찾는 감독 차원 감소법이다. 그러나 클래스 정보를 가진 데이터(labeled data, 라벨드 데이터)의 수가 매우 적을 때 LDA의 성능은 급격히 저하된다. 이러한 경우에 라벨드 데이터에 비해 수집이 상대적으로 쉬운 클래스 정보를 가지지 않은 데이터(unlabeled data, 언라벨드 데이터)를 활용하는 반감독(semi-supervised) 차원 감소 방법을 사용하여 학습 성능을 향상시킬 수 있다[1,2,3].

최근에 LDA를 확장한 반감독 차원 감소법인 ELDA/LNP(Extended LDA using LNP)[4]가 제안되었고, 다른 반감독 차원감소법과 비교하여 ELDA/LNP에 의해 차원이 감소되었을 때 분류성능이 크게 향상되는 것을 볼 수 있었다. ELDA/LNP는 전체 언라벨드 데이터 중에서 높은 신뢰도로 클래스 라벨이 예측될 수 있는 데이터를 선택하여 라벨드 데이터와 함께 LDA를 수행한다. 언라벨드 데이터의 예측

※ 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2011-0007779).
† 정 회 원 : 충남대학교 컴퓨터공학과 부교수
논문접수 : 2011년 10월 25일
수 정 일 : 1차 2012년 2월 1일
심사완료 : 2012년 2월 8일

신뢰도를 반영하기 위하여 그래프를 기반으로 한 확장된 LDA가 사용되었다. 그러나, ELDA/LNP는 다른 통계적 차원 감소법에서처럼 전체 데이터 샘플 수와 데이터 차원에 비례하여 연산 비용이 증가하는 역행렬연산이나 singular value decomposition(SVD) 등의 행렬 연산을 필요로 한다. 또한 LDA에서와 마찬가지로 얻어지는 특징값 차원의 수가 클래스의 수를 넘지 못하는 한계가 있다.

언라벨드 데이터를 모으는 것은 라벨드 데이터를 얻는 것에 비해서 상대적으로 비용이 적게 들지만 그렇다고 해서 가능한 많은 양의 언라벨드 데이터를 사용하는 것이 항상 도움이 될 것인가 하는 문제가 있다. 우리는 반감독 차원 감소에서 라벨드 데이터의 수가 적을 때 너무 많은 언라벨드 데이터로부터의 정보는 오히려 장애가 될 수 있다는 것을 보인다. 또한, 데이터 처리 속도와 메모리 사용 관점에서 다루기 용이한 수준을 유지하면서 많은 양의 언라벨드 데이터로부터 유용한 정보를 활용하기 위해 앙상블 방법을 이용한 반감독 차원 감소법을 제안한다. 2절에서 반감독 차원 감소법에 대해 간단히 소개하고, 3절에서 앙상블 기법을 이용한 반감독 차원 감소법을 제안한다. 4절에서 적은 수의 라벨드 데이터에 비해 많은 언라벨드 데이터의 수가 성능에 미치는 영향을 실험을 통해서 보이고 제안하는 앙상블 기법의 성능을 검증한다.

2. 반감독 차원 감소법

선형 차원 감소는 (1)과 같이 고차원공간에서 저차원공간으로 데이터를 매핑시키는 변환행렬 G로 나타낼 수 있다.

$$x \mapsto G^T x \tag{1}$$

선형판별분석법(Linear discrimiant analysis, LDA)는 클래스 간 거리를 최대화하고 클래스 내 분산을 최소화하는 변환행렬 G를 구함으로써 차원 감소 후 분류작업을 시행하는 것이 목적일 때 분류성능을 향상시킬 수 있게 한다.[5] 그러나 라벨드 데이터의 수가 매우 적을 때 LDA의 성능은 급격히 저하되게 된다. 이러한 경우에 손쉽게 얻을 수 있는 언라벨드 데이터로부터의 정보를 활용함으로써 적은 수의 라벨드 데이터의 한계를 극복하고자 하는 것이 반감독 학습법이다.

SSLDA[1]는 언라벨드 데이터를 포함한 모든 데이터 쌍들 사이의 유사도(similarity)를 나타내는 그래프를 이용하여 라벨드 데이터에 의한 클래스 내 분산의 최소화와 함께 유사도가 높은 임의의 데이터들 사이의 거리가 작게 되는 변환행렬을 구하였다. NSSLDA[3]는 클래스 간 거리를 최대화하는 데 언라벨드 데이터 정보를 반영하기 위하여 상이도(dissimilarity)가 큰 데이터들 사이의 거리가 멀리 떨어지도록 함으로써 SSLDA를 확장하였다. ELDA/LNP[4]는 전체 언라벨드 데이터 중에서 높은 신뢰도로 클래스 라벨이 예측될 수 있는 데이터를 선택하여 라벨드 데이터와 함께 LDA를 수행한다. 그러나 잘 못 예측된 라벨은 클래스 분포를 본래의 분포와 다르게 만들 수 있기 때문에 언라벨드 데이터를 포함할 때 예측에 대한 신뢰도가 반영되도록 그래프를 이용하여 확장된 LDA를 수행한다.

2.1 ELAD/LNP

$X=[x_1,\cdots,x_n]=[x_1^1,\cdots,x_{n_1}^1,\cdots,x_1^r,\cdots,x_{n_r}^r]$는 r개의 클래스 중의 하나에 속하는 라벨드 데이터들의 셑이고 $U=[x_{n+1},\cdots,x_{n+u}]$는 언라벨드 데이터 셑이라고 할 때, 반감독 분류기인 LNP(Linear neighborhood propagation)[6]에 의해 각 언라벨드 데이터 샘플에 대해 $[t_1,\cdots,t_r]$을 구한다. 여기서 t_i는 데이터샘플이 클래스 i에 속할 확률로 해석할 수 있다. 따라서 데이터샘플은 가장 큰 t_i값을 가지는 클래스로 예측되며 t_i값들로부터 데이터 샘플의 예측된 클래스라벨에 대한 신뢰도 $\max_{1\le i\le r} t_i / sum_{1\le i\le r} t_i$ 를 구한다. 모든 언라벨드 데이터 중에서 높은 신뢰도를 가진 것들을 예측된 클래스의 멤버로서 포함시킴으로써 증가된 수의 라벨드 데이터 셑 $X_*=[x_1,\cdots,x_{n^*}]=[x_1^1,\cdots,x_{n_1^*}^1,\cdots,x_1^r,\cdots,x_{n_r^*}^r]$를 얻을 수 있다. ELDA/LNP에서는 X_*에 대해 차원감소법 LDA를 수행하기 위해 신뢰도를 반영하는 유사도 그래프와 패널티 그래프의 가중치 행렬 $S=\{s_{ij}\}$와 $P=\{p_{ij}\}$을 다음과 같이 정의한다.

$$s_{ij}=\begin{cases}\frac{1}{n_k^*}\times b_i\times b_j & ;\ x_i\text{와 } x_j\text{가 클래스 } k\text{에 속할 때}\\ 0 & ;\text{그 밖의 경우}\end{cases}$$

$$p_{ij}=\frac{1}{n^*}\times b_i\times b_j - s_{ij}$$

여기서 b_i는 데이터샘플 x_i의 클래스라벨 예측에 대한 신뢰도를 나타낸다. 이제 다음을 만족하는 차원감소를 위한 정사영 벡터를 구할 수 있다.

$$\begin{aligned}\arg max_g \frac{\sum_{i=1}^{n^*}\sum_{j=1}^{n^*} p_{ij}(g^Tx_i - g^Tx_j)^2}{\sum_{i=1}^{n^*}\sum_{j=1}^{n^*} s_{ij}(g^Tx_i - g^Tx_j)^2}\\ = \arg max_g \frac{g^T X_*(D_p-P)X_*^T g}{g^T X_*(D_s-S)X_*^T g}\end{aligned}$$

D_s와 D_p는 i번째 성분이 각각 $\sum_j s_{ij}$와 $\sum_j p_{ij}$인 대각행렬이다.

$$X_*(D_p-P)X_*^T g = \lambda X_*(D_s-S)X_*^T g \tag{2}$$

의 가장 큰 고유값들에 대응되는 고유벡터들이 선형변환행렬 G를 구성하게 된다.

3. 앙상블 방법을 이용한 반감독 차원 감소

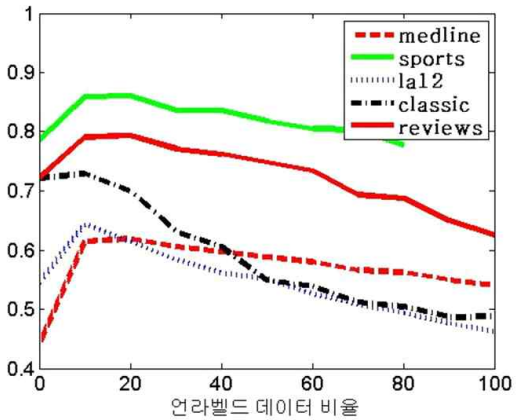
3.1 언라벨드 데이터의 사이즈가 반감독 차원감소에서 미치는 영향 분석

ELDA/LNP는 언라벨드 데이터의 라벨을 예측하고 신뢰도를 측정하기 위해 반감독학습법인 LNP를 이용하였다. LNP는 라벨드 데이터의 라벨정보를 이웃하고 있는 언라벨드 데이터에 전파하고 다시 이것을 그 주위의 다른 데이터로 전파함으로써 언라벨드 데이터의 라벨을 예측하는 방법이다. 따라서 데이터 샘플들의 가까운 이웃들을 탐색하는 과정과 역행렬연산이 필요하다. 이러한 연산들은 전체 데이터 샘플들의 수에 비례하여 증가하는 처리시간과 메모리 사용량을 요구한다. 그래프를 기반으로 하는 확장된 LDA 방법 또한 식 (2)에서와 같이 고유값과 고유벡터를 구하는 행렬연산이 필요하므로 데이터 수와 데이터 차원이 클수록 처리비용이 증가하게 된다[7].

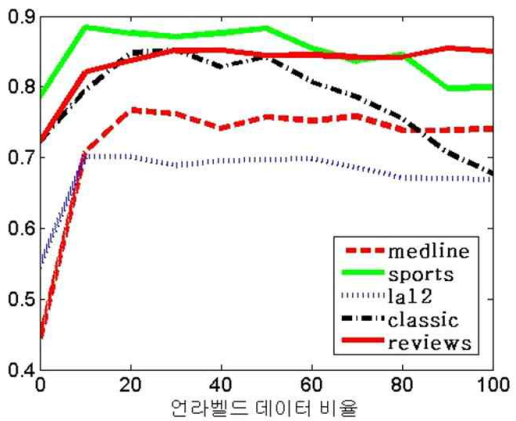
언라벨드 데이터를 얻는 비용은 적게 들지만 많은 양의 데이터를 모두 사용한다고 더 좋은 성능이 나오지는 않는다는 것을 보이기 위해 반감독 차원감소에서 언라벨드 데이터의 사이즈가 미치는 영향에 대해 살펴본다. 주요한 전제는 너무 많은 언라벨드 데이터로부터의 정보는 적은 수의 라벨드 데이터를 무기력하게 만든다는 것이다. 이것을 테스트하기 위해 <표 1>에 요약한 문서데이터를 이용하였다. medline은 [8]에서 사용된 의학논문의 초록들을 모은 데이터베이스에서 추출된 텍스트데이터이고 나머지 데이터는 [9]에서 다운받은 텍스트 데이터 모음 중에서 선택한 데이터이다. 각 데이터셋에 대해 30%를 테스트 데이터로 사용하였다. 70%의 트레이닝 데이터중에서 클래스 당 5개의 원소를 라벨드 데이터로 사용하고 나머지에서 수를 증가시켜가면서 언라벨드 데이터의 사이즈를 변화시켰다. 라벨드 데이터와 언라벨드 데이터에 대해 반감독 차원 감소법을 적용하여 차원감소 변환행렬을 구하고 1-nearest neighbor classifier를 이용하여 테스트 데이터의 분류정확도를 측정하였다. 테스트와 트레이닝 데이터 분리를 랜덤하게 10번 반복하여 평균 분류정확도를 측정하여 (그림 1)과 (그림 2)에 나타내었다. x축은 70%의 트레이닝 데이터 중에서 언라벨드 데이터로 사용된 데이터의 비율을 나타내며 10%에서 100%까지 변화시켜가며 측정하였다. (그림 1)은 SSLDA에 대한 실험 결과이고 (그림 2)는 ELDA/ LNP를 실행한 결과이다. 두 그래프는 언라벨드 데이터의 사이즈가 커질수록 반감독 차원 감소의 성능이 향상되지는 않는다는 것을 보여준다. 오히려 10%-30%내외의 범위에서 높은 분류정확도를 얻을 수 있었다.

〈표 1〉 데이터 셀 설명

데이터셀	다큐먼트 수	클래스 수	속성 수
medline	2500	5	22095
sports	8313	5	18136
la12	6279	6	21604
classic	7094	4	12009
review	3932	4	23165



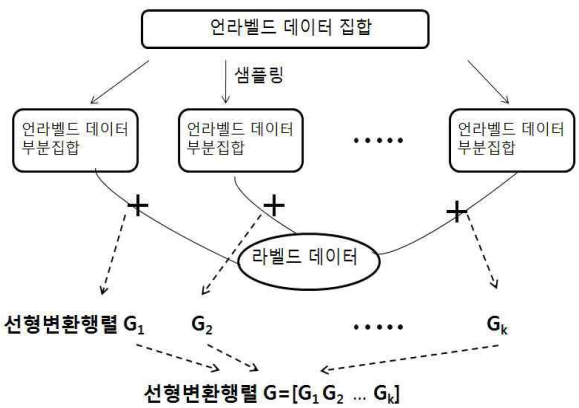
(그림 1) SSLDA에서 언라벨드 데이터의 사이즈의 변화에 따른 분류정확도



(그림 2) ELDA/LNP에서 언라벨드 데이터의 사이즈의 변화에 따른 분류정확도

3.2 반감독 차원 감소를 위한 앙상블 방법

우리가 제안하는 앙상블 방법은 감독학습에서의 앙상블 방법의 변형된 버전이다. 감독학습에서는 라벨드 트레이닝 셀에서 샘플링 된 데이터로 학습된 분류기를 다수 구성하고 이것들을 결합하여 강한 분류기를 만들어낸다. 반감독 차원 감소를 위한 앙상블 방법에서는 소수의 라벨드 데이터는 항



(그림 3) 앙상블 방법을 이용한 반감독 차원 감소 방법 개요

상 포함시키면서 샘플링은 언라벨드 데이터 셀에서 이루어진다. 이제 샘플링 된 데이터에 대해 ELDA/LNP를 수행하여 선형변환행렬 G_i를 구한다. 이 과정을 반복하면서 얻어진 선형변환 행렬들을 $G=[G_1, G_2, \dots, G_k]$로 연결하여 많은 양의 언라벨드 데이터를 효과적으로 활용하는 차원감소 변환을 구할 수 있다. 이 과정을 (그림 3)에 나타내었다.

4. 실험 및 결과 분석

제안한 앙상블 방법을 테스트하기 위해 3.1절에서와 같은 실험환경에서 언라벨드 데이터의 크기를 클래스의 수*100으로 고정하여 진행하였다. 앙상블 멤버의 수는 10개로 하여 변환행렬 $G=[G_1, \dots, G_{10}]$을 구하였다. 따라서 <표 1>의 각 데이터 셀에 대해 데이터 차원은 10*(클래스의 수-1)로 감소된다. 이 실험을 20번 반복하여 평균 분류정확도를 <표 2>에 나타내었다. <표 2>의 두 번째 칼럼은 라벨드 데이터만을 사용하여 LDA를 수행한 결과이고 세 번째와 네 번째 칼럼은 사용가능한 모든 언라벨드 데이터를 사용하여 LNP와 ELDA/LNP를 수행한 결과이다. 다섯 번째 칼럼은 제안한 앙상블 방법에 의해 얻어진 결과를 나타낸다. 표에서 보여주는 것처럼 앙상블 방법을 이용한 ELDA/LNP는 거의 모든 데이터 셀에서 나은 성능을 보여준다.

〈표 2〉 앙상블 ELDA/LNP의 성능 비교

데이터	LDA	모든 언라벨드 데이터 사용		앙상블
		LNP	ELDA/LNP	ELDA/LNP
medline	0.439	0.666	0.740	0.762
sports	0.786	0.744	0.799	0.912
la12	0.545	0.579	0.669	0.734
classic	0.722	0.535	0.678	0.828
review	0.722	0.822	0.850	0.852

5. 결 론

우리는 이 논문에서 적은 수의 라벨드 데이터와 함께 많은 언라벨드 데이터를 가지고 있을 때 대용량 데이터 처리 문제를 극복하면서 동시에 언라벨드 데이터의 정보를 효과적으로 사용할 수 있도록 앙상블 방법을 이용한 반감독 차원 감소법을 제안하였다. 각 앙상블 멤버에 대해 사이즈가 작은 언라벨드 데이터 부분집합을 이용하여 차원 감소를 수행하므로 수행시간을 절약할 수 있으면서 동시에 앙상블 방법으로 모든 언라벨드 데이터가 골고루 이용될 수 있게 된다. 특히 제안된 방법은 배깅(bagging)에 의한 앙상블 방법을 이용함으로써 변환행렬 G_i를 구하는 과정에 대해 병렬연산을 수행할 수 있다는 장점이 있다. 이러한 앙상블 방법은 다른 반감독 학습법에서도 적용할 수 있을 것이라 기대된다.

참 고 문 헌

[1] Y. Song, F. Nie, C. Zhang, and S. Xiang, “A unified framework for semi-supervised dimensionality reduction,” Pattern recognition, Vol.41, pp.2789-2799, 2008.

[2] D. Cai, X. He, and J. Han, “Semi-supervised discriminant analysis,” Proceedings of the international conference on computer vision, 2007.

[3] G. Lim and C. H. Park, “Semi-supervised dimension reduction using graph-based discriminant analysis,” Proceedings of the international conference on computer and information technology, 2009.

[4] Y. Lee, Y. Shin, and C. H. Park, “Extending linear discriminant analysis by using unlabeled data,” Proceedings of the international conference on computer and information technology, 2011.

[5] K. Fukunaga, “Introduction to Statistical Pattern Recognition,” second edition, Academic Press, 1990.

[6] J. Wang, F. Wang, C. Zhang, H. Shen, and L. Quan, “Linear neighborhood propagation and its application,” IEEE transactions on pattern analysis and machine intelligence, Vol.31, No.9, pp.1600-1615, 2009.

[7] G. H. Golub and C. F. Loan, “Matrix computations,” 3rd edition, The Johns Hopkins University Press, 1996.

[8] H. Kim, P. Holland and H. Park, “Dimension reduction in text classification with support vector machines,” Journal of machine learning research, Vol.6, pp.37-53, 2005.

[9] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download

박 정 희

e-mail : cheonghee@cnu.ac.kr

1991년 연세대학교 수학과(학사)

1998년 연세대학교 수학과(이학박사)

2004년 미네소타대학교 컴퓨터공학과 (공학박사)

2005년~2006년 충남대학교 컴퓨터공학과 전임강사

2007년~2010년 충남대학교 컴퓨터공학과 조교수

2011년~현 재 충남대학교 컴퓨터공학과 부교수

관심분야 : 데이터마이닝, 패턴인식 등