다변량 데이터 분석을 위한 PCA 알고리즘 구현

김귀숙*, 손호선*, 류근호*, 이영성*
*충북대학교 의생명과학경영융합 대학원
e-mail : kksuk@dblab.chungbuk.ac.kr

Performance of PCA Algorithm for Multivariate Data Analysis

GwiSuk Gim *, Ho Sun Shon*, Keun Ho Ryu*, YoungSung Lee*
*Graduate School of Health Science Business Convergence, Chungbuk National University

요 약

다변량 데이터 분석에 주로 사용되는 차원축소 기법 중 하나인 PCA 알고리즘을 직접 구현해보고 기존의 통계분석 프로그램과 그 결과를 비교분석 해보았다. UCI 에서 제공하는 유방암 데이터를 이용하여 실험 해본 결과 두 프로그램 모두 같은 주성분을 얻고, Eigenvalue 와 variance 도 같은 값을 얻었다. 따라서 상용화된 통계패키지를 사용하지 않고도 PCA 알고리즘을 적용하여 차원축소 문제를 해결하고 데이터를 분석 할 수 있다.

1. 서론

최근 데이터들은 데이터의 수집과 저장 기술의 급속한 발전으로 인해 점점 그 크기가 거대하게 변하고 있다. 이러한 대규모 데이터를 분석하기 위하여 데이터 마이닝 기법과 다양한 통계적 분석 방법에 대한 요구 또한 커지고 있다.

특히, 분석하고자 하는 데이터의 속성이 점점 증가하면서 발생하는 문제인 차원의 저주 문제를 해결 하는 것이 필요하다.

차원의 저주란 속성의 수가 많아지면서 오히려 성능에 나쁜 영향을 미쳐 분류기에 의한 학습과 인식 속도가 더 느려진다는 의미이다. [1]

이러한 차원의 저주 문제를 해결하기 위해서는 고차원의 데이터를 저차원의 데이터로 변환시켜 주어야 한다. 이러한 차원 축소에 사용하는 알고리즘으로 PCA(Principal Component Analysis), LDA(Linear Discriminant Analysis), ICA(Independent Component Analysis) 등이 있다.

이중 가장 널리 쓰이는 알고리즘은 PCA 로 PCA 알고리즘을 이용하여 데이터를 분석하기 위해서는 기존의 통계 패키지인 SPSS 나 R 을 사용한다. 이와 같은 통계 패키지는 다양한 통계 분석 방법을 제공해주지만 단점이 존재한다. SPSS 의 경우 제품 라이선스가 비싸고, R 의 경우 데이터를 분석하기 위하여 계산 함수를 이용하여 프로그래밍을 할 필요가 있다.

본 연구의 목적은 위와 같은 기존의 통계 패키지의 단점을 보완하고자 User-Interface 중심의 PCA 알고리즘을 직접 구현하고 그 결과를 시각화하여 쉽게 데이터의 분석을 할 수 있도록 하는 것이다. 구현한 PCA 알고리즘은 다변량 데이터를 분석하여 기존의 통계 패키지와 그 결과를 비교분석하여 그 성능을 평가하였다.

2. 관련 연구

PCA(Principal Component Analysis)는 데이터를 고차원 공간에서 저차원 공간으로 투영하기 위한 선형 대수 기법으로 데이터의 차원축소에 주로 사용하는 분석 방법 중 하나이다.[1] PCA 를 활용하면 관측대상이 어떠한 위치에 있는지 시각적으로 파악할 수 있게 된다.[3]

이러한 PCA 알고리즘을 사용하기 위해서는 주로 기존의 상용화 되어있는 통계패키지를 이용한다. 이러한 프로그램은 SPSS, R 등이 있다.

SPSS 는 1968 년 SPSS Inc.에 의해 제작 되었으며, 2013 년 현재 IBM 에 인수되어 IBM SPSS Statistics 22.0 버전이 발매되어있다. SPSS 는 통계 분석에서 가장 널리 사용되는 프로그램 중 하나로써 선형 및 비선형 모델링, 상관분석, 회귀분석, 비모수 검정, 시뮬레이션 기능 등 다양한 통계 분석 방법을 제공한다.[4] 하지만, SPSS 는 제품 라이선스가 매우 비싸고, 많은 통계 분석 방법과 기능을 제공하기 때문에 사용하고자 하는 분석방법과 프로그램 사용에 대한 학습이 필요하며 데이터의 시각화에 취약한 것이 단점이다.

R 은 통계 계산 및 그래픽을 위한 무료 소프트웨어 프로그래밍 언어이다. R 은 선형 및 비선형 모델링, 고전적 통계검사, 시계열 분석, 분류, 클러스터링 등을 제공한다.[5] R 은 무료 소프트웨어로 소스코드가 공개되어 있기 때문에 사용자가 직접 새로운 기능이나 알고리즘을 제작하여 패키지 형태로 추가할 수 있다. R 은 matlab 과 같은 언어 중심 프로그램이기 때문에 데이터를 분석하기 위해서는 R 언어와 분석에 사용하는 여러 함수들을 학습하여 사용자가 직접 프로그래밍을

해야하는 단점이 있다.

3. PCA 알고리즘 구현

PCA 알고리즘은 matlab 언어를 이용하여 Stand alone 형식으로 구현하였다. 구현한 PCA 알고리즘을 이용하여 데이터를 분석하려면 먼저, 분석 하고자 하는 데이터를 엑셀(.xls) 형식 파일로 저장한 뒤 프로그램의 File 메뉴의 데이터 입력 기능을 이용하여 데이터를 입력해준다. bilud model 메뉴의 PCA 기능을 사용하여 데이터를 분석한 뒤 같은 메뉴의 Score Plot, Loading Plot 기능으로 데이터 분석 결과를 시각화하여 나타낸다.

PCA 알고리즘의 계산방법에는 공분산행렬(Covariance matrix)을 이용하는 방법과 상관행렬(Correlation matrix)을 이용하는 방법 두 가지가 있다.

이 프로그램에서는 상관행렬을 이용하여 주성분을 계산한다. 아래는 프로그램에 사용한 PCA 의 계산 과정이다. X(m×n 행렬)는 입력 데이터, X^T는 X 의 전치행렬이다. 상관행렬 R 은 다음과 같이 정의된다.

$$R = \frac{X^T \times X}{(m-1)} \quad (1)$$

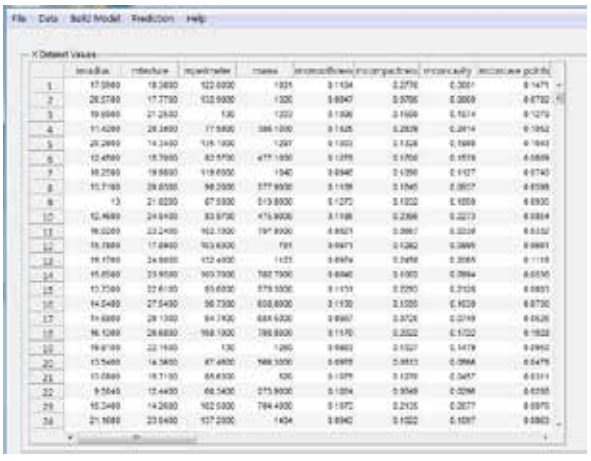
$$R = USV^T \quad (2)$$

(1)의 식에서 구한 상관행렬을 SVD(Singular Value Decomposition)[4]를 이용하여 U (m × m 직교행렬), S(m×n 대각행렬), V^T(n×n V의 전치행렬로 unitary 행렬)를 계산한다. 이 때, 행렬 S는 데이터의 Eigenvalue 를 포함하고, 행렬 V^T는 Eigenvector 를 포함하고 있다.

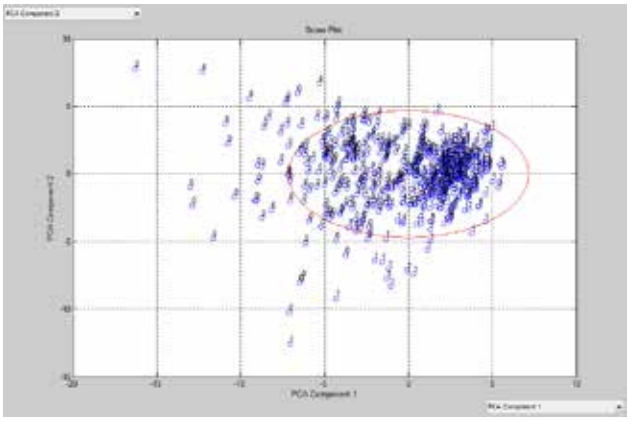
$$P = V^T \quad (3)$$

$$T = X \times P \quad (4)$$

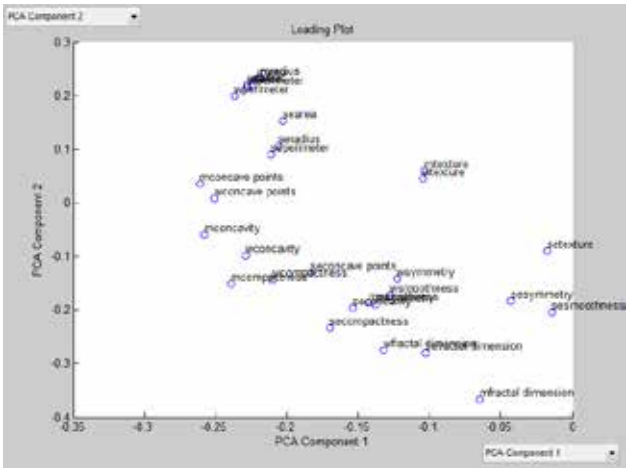
(3)의 식에서 P는 Loading 행렬, (4)의 식에서 T는 Score 행렬이다. Loading 행렬은 Loading Plot 을 그리는데 사용되며 각 변수들의 상호관련성을 보여준다. Score 행렬로는 Score Plot 을 그릴 수 있으며 전체 데이터 값들의 분포와 특징을 쉽게 분석할 수 있다.



(그림 1) PCA 알고리즘을 적용한 데이터 입력 형태



(그림 2) PCA 알고리즘을 적용한 Score Plot



(그림 3) PCA 알고리즘을 사용한 Loading Plot

4. 실험 및 결과

실험에 사용된 데이터는 UCI[7]에서 제공하는 유방암 데이터, 와인 데이터, 이스트 데이터이다. 유방암 데이터는 샘플 569 개, 속성 32 개로 이루어져있고 이 중 ID number 와 Diagnosis 2 가지를 제외한 30 개의 속성을 가지고 실험을 진행하였다. 와인 데이터는 샘플 4898 개, 속성 12 개, 이스트 데이터는 샘플 1484 개, 속성 8 개로 이루어져있다.

분석 도구는 직접 구현한 PCA 알고리즘과 SPSS 의 PASW Statistics 18.0.0, R 3.0.1 에서 패키지 vegan 을 사용하였다.

아래의 표에서 볼 수 있듯이 구현한 PCA 알고리즘과 기존 통계프로그램의 PCA 알고리즘 계산 결과는 소수점 아래 자릿수 표시로 인한 반올림의 차만 있을 뿐 같은 결과값을 나타내는 것을 알 수 있다.

<표 1> 구현한 PCA 알고리즘, SPSS, R 의 Wisconsin Diagnostic Breast Cancer data 를 이용한 주성분 계산 결과 비교

Principle	구현한 PCA 알고리즘	SPSS	R

Component	Eigenvalue		
1	13.2816	13.282	13.2816
2	5.6914	5.691	5.6914
3	2.8179	2.818	2.8179
4	1.9806	1.981	1.9806
5	1.6487	1.649	1.6487
6	1.2074	1.207	1.2074

<표 2> 구현한 PCA 알고리즘, SPSS, R 의 Wine Quality_red Data 를 이용한 주성분 계산 결과 비교

Principle Component	구현한 PCA 알고리즘	SPSS	R
	Eigenvalue		
1	3.1212	3.121	3.12117
2	2.2419	2.242	2.24188
3	1.6829	1.683	1.68292
4	1.2150	1.215	1.21502

<표 3> 구현한 PCA 알고리즘, SPSS, R 의 Yeast Data 를 이용한 주성분 계산 결과 비교

Principle Component	구현한 PCA 알고리즘	SPSS	R
	Eigenvalue		
1	1.8142	1.814	1.8142
2	1.2703	1.270	1.2703
3	1.0212	1.021	1.0212

5. 결론

본 논문에서는 다변량 데이터 분석 방법 중 하나인 PCA 알고리즘을 구현하고 기존의 통계 프로그램과 비교 분석하였다. 또한 Score Plot 과 Loading Plot 을 시각화 하였다. 그 결과 기존 프로그램과 같은 결과를 보였으므로 이를 활용하여 다변량 데이터 분석을 할 수 있다.

향후 연구로는 다양한 다변량 분석방법을 구현하고, 다른 알고리즘과의 앙상블을 통해 데이터 분석 성능을 높이는 연구를 진행할 것이다. 또한 PCA 를 통해 얻은 결과값에 Cross validation 을 적용시켜 학습모델을 만들어 분류가 가능하도록 할 것이다.

사사 표기

본 논문은 2013 년도 미래창조과학부의 재원으로 과학벨트기능지구지원사업의 지원을 받아 수행된 연구임(2013K001552).

참고문헌

[1] Tan el.al. “Introduction to Datamining”

[2] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Morgan kaufmann, 2012.

[3] 노형진, “SPSS 에 의한 다변량데이터의 통계분석”, 도처출판 효산, 2007

[4] http://www.ibm.com/kr/ko/

[5] http://www.r-project.org/

[6] Demmel, James W. “Applied numerical linear algebra”. Siam, 1997.

[7] http://archive.ics.uci.edu/ml/index.html