Journal of the Korean Institute of Industrial Engineers
Vol. 40, No. 3, pp. 275-282, June 2014.
ISSN 1225-0988 | EISSN 2234-6457
http://dx.doi.org/10.7232/JKIIE.2014.40.3.275

<Original Research Paper>

주성분 분석 로딩 벡터 기반 비지도 변수 선택 기법

박영준 · 김성범†

고려대학교 산업경영공학과

Unsupervised Feature Selection Method Based on Principal Component Loading Vectors

Young Joon Park · Seoung Bum Kim

School of Industrial Management Engineering, Korea University

One of the most widely used methods for dimensionality reduction is principal component analysis (PCA). However, the reduced dimensions from PCA do not provide a clear interpretation with respect to the original features because they are linear combinations of a large number of original features. This interpretation problem can be overcome by feature selection approaches that identifying the best subset of given features. In this study, we propose an unsupervised feature selection method based on the geometrical information of PCA loading vectors. Experimental results from a simulation study demonstrated the efficiency and usefulness of the proposed method.

Keywords: High-Dimensional Data, Unsupervised Feature Selection, Filter Method, Principal Component Analysis

1. 서 론

최근 데이터 수집 기술과 저장 기술의 발달로 대용량 고차원 데이터가 도처에서 생성되고 있으며 이에 대한 효과적인 분석을 위해 차원축소는 중요한 연구주제로 대두되고 있다(Kim *et al*., 2011). 특히, 관측치의 수 보다 변수의 수가 많은 경우 계산 자체가 불가능한 경우가 많으며, 예측 정확도, 계산 비용, 그리고 시각화의 부분에서 분석의 어려움을 야기할 수 있다(Malhi *et al*., 2004).

차원 축소 기법은 크게 변수 추출 기법과 변수 선택 기법으로 나눌 수 있다(Mao, 2005). 변수 추출 기법은 기존 변수의 결합을 통해 새로운 변수를 생성하는 방법이다(Guyon *et al*., 2003). 따라서 추출된 변수는 원 변수 간의 상관관계를 반영하고 있다. 하지만 생성된 변수는 고차원 원 변수들의 결합으로 정의되기 때문에 추출변수에 대한 해석이 어렵다는 한계가 있다. 변수 추출 기법은 반응 변수의 존재 유무에 따라 지도 변수 추출 기법과 비지도 변수 추출 기법으로 나눌 수 있다. 반응 변수가 존재하는 경우 대표적인 기법으로는 PLS(partial least square)가 있으며 비지도 변수 추출 기법의 대표적인 기법으로는 주성분 분석이 있다(Widjaja *et al*., 2012; Kim, 2009).

차원 축소의 또 다른 방법으로는 변수 선택기법이 있는데 이는 원 변수에서 주요 변수를 선정하는 것이다. 이렇듯 변수 선택 기법은 원 변수의 변환을 거치지 않기 때문에 변수 추출 기법에서 야기되었던 해석 문제를 해결할 수 있다. 하지만, 대부분의 기법이 변수 선택 시 변수간의 상관관계를 고려하지 않는다는 한계가 있다(Guyon *et al*., 2003). 변수 선택 기법 역시 반응 변수의 사용 유무에 따라 지도 변수 선택 기법과 비지도 변수 선택 기법으로 나눌 수 있다. 지도 변수 선택법은 비교적 많은 연구가 행해져 왔으며 대표적으로는 회귀분석을 이용한 전향/후향/계단식 기법이다. 비지도 변수 선택법은 반응 변수

본 연구는 미래창조과학부의 재원으로 한국연구재단의 기초연구사업(2013007724)과 지식경제부 정보 통신 기반 구축 사업의(NIPA-2011-(B1110-1101-0002)) 지원을 받아 수행됨.
† 연락저자 : 김성범 교수, 136-701 서울시 성북구 안암동 5가 1번지 고려대학교 산업경영공학과, Tel : 02-929-5888, Fax : 02-3290-3397, E-mail : sbkim1@korea.ac.kr
2013년 12월 26일 접수; 2014년 3월 19일 수정본 접수; 2014년 5월 17일 게재 확정.

가 없는 경우 주요 변수를 찾는 방법이다. 이 기법의 관한 연구는 군집화 기법에 특성화 된 변수 선택 기법이 일부 행해졌으나 다른 차원 축소 기법에 비해 연구가 덜 진행된 분야이다(Dash *et al*., 2002; Roth *et al*., 2003; Guyon *et al*., 2003, Mao, 2005). 일반적으로 변수 선택 기법은 변수 추출 기법에 비해 차원 축소의 정도가 적지만 해석상의 이점이 있기 때문에 해석이 중요한 문제에서 널리 사용되고 있다. 특히, 비지도 변수 선택 기법은 지도 변수 선택 기법에 비해 과적합 경향이 작은 장점이 있다(Guyon *et al*., 2003).

제안 기법은 차원 축소 기법 중의 하나인 주성분 분석을 이용하였는데 이를 이용하여 축소된 차원은 원 변수의 선형 결합으로 표현되기 때문에 변수 선택의 중요한 정보로 사용할 수 있다(Guo *et al*., 2002). 이러한 점에 착안하여 주성분 분석을 이용한 변수 선택 기법에 관한 연구가 진행되었다. Jolliffe는 주성분 분석을 이용하여 고유값 상위 *k*개에 대응하는 추출된 변수를 기준으로, 이들 추출된 변수와 상관계수가 높은 변수를 선택하고 그렇지 않은 변수를 제거하는 방법을 제안하였다(Jolliffe, 1972). 이 기법은 매우 간단하지만 추출된 *k*개에 변수만을 고려하기 때문에 제한된 정보를 바탕으로 변수를 선택한다는 한계가 존재한다. 또 다른 기법으로는 주성분 분석과 이동평균 관리도 기법을 이용한 비지도 변수 선택 기법이(Kim *et al*., 2011). 이 방법은 주성분 분석으로부터 얻은 로딩을 모두 이용하여 변수의 중요도를 계산하였는데, 로딩값 모두가 변수의 중요도에 영향을 미치지 않을 수 있는 상황을 고려하지 않은 한계가 있다. 이와 관련한 자세한 내용은 본 논문의 제 2장에 기술하였다.

본 연구에서 제안하고 기법은 주성분 분석으로부터 얻은 로딩벡터의 기하학적 정보를 기반으로 주요 변수를 선택한다. 이러한 접근은 기존의 연구와 달리 주성분 분석의 로딩의 특성을 더 정교하게 반영하여 중요한 변수를 선택할 수 있도록 한다.

이후 본 논문의 구성은 다음과 같다. 본 논문의 제 2장에서는 주성분 분석을 이용한 제안하는 기법에 대해 설명했다. 제 3장과 제 4장에는 시뮬레이션과 실제 데이터를 이용한 실험을 통해 제안하는 기법을 검증하였으며, 제 5장에서는 결과와 앞으로의 연구 방향을 기록하였다.

2. 제안 방법

주성분 분석은 고차원 데이터의 차원을 축소 하기 위한 방법이다(Boutsidis *et al*., 2008). 주성분 분석으로부터 추출된 변수는 기존 변수들의 선형결합으로 구성되며 소수의 추출 변수로써 데이터를 설명할 수 있기 때문에 고차원 데이터 차원 축소에 효과적으로 활용 되고 있다(Wang *et al*., 2014). 이때 추출된 변수는 데이터의 공분산 행렬로부터 얻어지는 고유값과 고유벡터를 통해 계산되며, 데이터를 설명하는 공간의 기저를 식 (1)로 표현한 재구성 오차(reconstruction error)를 최소화 하도록 결정한다(Hastie *et al*., 2001).

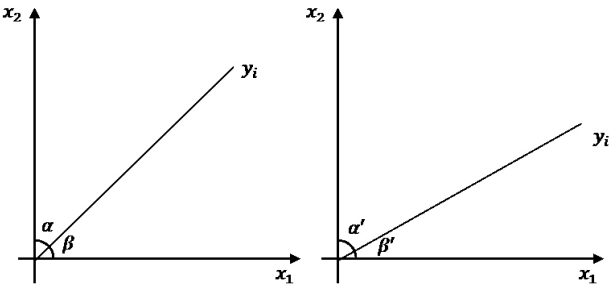

Figure 1. An interpretation of the angle between an extracted feature and original features

$$min_{\mu,\{\lambda_i\},V_q}\sum_{i=1}^{N} \| x_i-(\mu+V_q\lambda_i) \|^2, \tag{1}$$

위 식에서 x_i는 관측치를 의미하며, $\mu+V_q\lambda_i$는 주성분 분석에 의해 새롭게 구성되는 공간의 기저를 나타낸다. 예를 들어, 데이터가 *p*개의 변수를 갖고 있다면, 주성분 분석으로부터 추출된 변수는 아래와 같이 나타낼 수 있다.

$$\begin{aligned} \mathrm{Y}_1 &= \alpha_{11}X_1+\alpha_{12}X_2+\cdots+\alpha_{1p}X_p, \\ \mathrm{Y}_2 &= \alpha_{21}X_1+\alpha_{22}X_2+\cdots+\alpha_{2p}X_p, \\ &\vdots \\ \mathrm{Y}_\mathrm{p} &= \alpha_{p1}X_1+\alpha_{p2}X_2+\cdots+\alpha_{pp}X_p, \\ &(\text{단, } \alpha_{ij},\ i,\ j=1,\ 2,\ \cdots,\ p). \end{aligned} \tag{2}$$

위 식에서 Y는 추출된 변수를 나타내며, X의 선형결합의 계수값인 로딩 α는 기존 변수가 추출된 변수를 구성하는데 기여하는 정도를 나타낸다. 제안 방법은 주성분 분석으로부터 얻은 로딩 벡터를 기반으로 다음의 두 과정을 거쳐 중요변수를 선택한다.

3.1 변수 중요도 산출 과정

주성분 분석을 데이터에 적용하면 추출된 변수를 이용하여 데이터를 재표현 할 수 있다. 추출된 변수의 로딩은 기존 변수 각각이 갖고 있는 분산 정보의 양을 의미한다. 예를 들면, <Figure 1>의 추출된 변수 y_i는 변수 x_1과 x_2의 정보가 사영되는 공간을 의미하며 $y_i=\cos\alpha\, x_1+\cos\beta\, x_2$로 표현할 수 있다. 만약 y_i가 x_1, x_2와 이루는 각도가 같다면(<Figure 1>(a)), 즉 $\alpha=\beta$인 경우에는 계수 $\cos\alpha$와 $\cos\beta$의 값이 같기 때문에 두 변수는 동일한 양의 분산정보를 갖는다고 해석할 수 있다. 반면, <Figure 2>(b)에서 보여주듯 y_i가 x_1, x_2와 이루는 각도가 다른 경우에는 $(\alpha'>\beta')$인 경우에는 각각의 $\cos\alpha'$와 $\cos\beta'$에 비례하는 분산 정보를 갖는다고 할 수 있다(<Figure 1>(b)).

만약 추출된 변수가 1사분면 이외의 공간에 위치할 경우, 즉 임의의 로딩이 음수 값을 갖는 경우에도 그 절대값이 크다면 많은 분산 정보를 포함하고 있음으로 음의 로딩 값에 대해서는 절대값을 취하여 고려한다.

결론적으로 주성분분석으로부터 추출된 변수는 다음과 같이 요약할 수 있다. 추출된 변수를 구성하는 로딩은 각 변수가 갖고 있는 분산정보를 내포하고 있으며, 만약 모든 변수가 동일한 분산 정보를 갖고 있다면 모든 로딩은 같은 값을 같게 되며 따라서 특별히 중요한 변수는 존재하지 않는 경우이다. 따라서 본 연구에서는 모든 변수가 동일한 중요도를 가질 때를 가정한 경우의 로딩보다 큰 로딩을 갖는 변수를 선택하도록 변수 선택 기법을 개발하였다.

좀 더 자세히 설명하기 위해 j번째 변수의 로딩 값을 갖고 있는 벡터를 다음과 같이 표현해 보자.

$$\boldsymbol{v}_j = [\alpha_{1j}, \alpha_{2j}, \cdots, \alpha_{pj}]^T \text{ for } j = 1, \cdots, p. \tag{3}$$

즉, 위 식에서 벡터 $\boldsymbol{v}_j$의 원소는 각 추출 변수를 구성하는 j번째 로딩 값이며 이는 해당 변수의 중요도를 의미한다. 여기서 좀 더 고려해야 할 사항은 추출 변수의 중요도가 다르기 때문에 이에 대한 가중치를 고려해야 하며 가중치는 각 추출변수의 분산 정보를 담고 있는 고유값(상관계수 행렬로부터 구한)을 이용할 수 있다. 따라서, 가중치가 고려된 로딩 벡터는 다음과 같다.

$$\tilde{\boldsymbol{v}}_j = \left[\tilde{v}_1\alpha_{1j}, \tilde{v}_2\alpha_{2j}, \cdots, \tilde{v}_p\alpha_{pj}\right]^T, \quad \text{where } \tilde{\lambda}_i = \frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i} \tag{4}$$

제안 기법의 핵심은 각 변수의 중요도를 살펴보기 위해 위 $\tilde{\boldsymbol{v}}_j$와 모든 변수가 동일한 중요도를 가질 때의 벡터 $\tilde{\boldsymbol{v}}_{ref}$를 구하여 두 벡터를 비교하는 것이다. 모든 변수가 동일한 중요도를 갖는다는 것은 각 변수의 선형결합을 구성하는 계수의 값이 같다는 것을 의미하며 기하학적으로는 변수가 이루는 공간에서 추출된 변수의 방향이 어느 한쪽으로 치우쳐져 있지 않고 모든 변수와 같은 각을 유지하는 방향을 갖는 것을 의미하며 다음과 같이 표현할 수 있다.

$$\tilde{\boldsymbol{v}}_{ref} = \left[\frac{1}{\sqrt{p}}, \frac{1}{\sqrt{p}}, \cdots, \frac{1}{\sqrt{p}}\right]^{\mathrm{T}} \tag{5}$$

위 벡터는 다음의 유도과정을 통해 설명할 수 있다. 모든 변수가 갖는 분산 정보의 양이 같은 경우 주성분 분석을 통해 임의의 추출된 변수는 다음과 같이 표현 가능하다.

$$Y = \sum_{i=1}^{p} \alpha\, \hat{e}_i. \tag{6}$$

위 식에서 $\hat{e}_i$는 i번째 변수에 해당하는 카테시안 좌표계의 기저벡터이다. Y는 다른 임의의 변수와 이루는 각이 동일 하기 때문에 임의의 기저벡터 $\hat{e}_k$와 내적을 통해 각을 계산할 수 있다.

$$\begin{aligned} \mathrm{Y} \cdot \hat{e}_k &= \| \mathrm{Y} \| \, \| \hat{e}_k \| \cos\theta \\ &= \sqrt{\textstyle\sum_{i=1}^{p}\alpha^2} \times \cos\theta \\ &= \alpha \end{aligned} \tag{7}$$

$$\therefore \cos\theta = \frac{\alpha}{\sqrt{\alpha^2 \sum_{i=1}^{p} 1}} = \frac{1}{\sqrt{p}} \tag{8}$$

$\cos\theta$는 추출된 변수 Y에서 k번째 변수에 해당하는 로딩을 의미한다. 식 (5)는 모든 변수가 동일한 분산 정보를 갖는 경우 기대되는 고유벡터이며 모든 원소가 동일한 값을 갖는 p차원의 $\mathbf{1} = [1, 1, \cdots, 1]^{\mathrm{T}}$ 벡터를 정규화한 벡터와 같다.

앞에서 구한 $\tilde{\boldsymbol{v}}_j$와 비교를 위해 $\boldsymbol{v}_{ref}$를 다음과 같이 가중치를 고려한 형태로 표현할 수 있다.

$$\tilde{\boldsymbol{v}}_{\mathrm{ref}} = \left[\tilde{\lambda}_1\frac{1}{\sqrt{p}}, \tilde{\lambda}_p\frac{1}{\sqrt{p}}, \cdots, \tilde{\lambda}_p\frac{1}{\sqrt{p}}\right]^T. \tag{9}$$

$\tilde{\boldsymbol{v}}_{\mathrm{ref}}$는 모든 변수가 동일한 중요도를 가진다는 가정 하에 임의의 변수가 가지는 로딩이므로 주성분 분석을 수행하여 얻은 실제 로딩 값과의 비교를 통해 각 변수가 얼마나 중요한지 결정할 수 있다. 변수의 중요도(I)는 두 벡터 $\tilde{\boldsymbol{v}}_j$, $\tilde{\boldsymbol{v}}_{\mathrm{ref}}$의 차를 이용하여 다음과 같이 정의한다.

$$\mathrm{I}(x_j) = \sum_{i=1}^{p} \tilde{\lambda}_i \psi\left(|\alpha_{ij}| - \frac{1}{\sqrt{p}}\right), \quad \text{where } \psi(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0. \end{cases} \tag{10}$$

식 (10)에서 $\psi(x)$는 중요도 함수에서 음수값을 제외하는 역할을 하는 함수다. 이것은 주성분 분석을 통해 구축되는 새로운 좌표계에 각 변수가 평균 이상의 정보를 사영할 때만 중요도로 반영하기 위해서이다. 만약 이런 과정이 중요도를 계산하는 것에서 제외된다면 기하학적 관점에서 문제가 야기 될 수 있는데 예를 들면, 어떤 변수가 주성분 분석에 의해 추출된 임의의 한 축과 이루는 각이 작아 중요도가 높았다면, 나머지 추출된 축과는 거리가 멀어질 가능성이 높아진다. 이런 상황에서는 중요도가 서로 상쇄되어 주요 변수를 감지하지 못할 수 있다.

3.2 변수 선택 기준

제 3.1절에서 제안한 방법으로 변수의 중요도에 순위를 정

했다면 이제는 변수를 선택하는 기준이 필요하다. 이를 위해 제안 방법은 선택할 변수의 수를 정하는 기준으로 식 (8)의 목적식을 이용한다. 목적식은 선택된 변수로 구성된 데이터의 분산정보량을 최대화 하는 동시에 선택된 변수의 수는 최소화 할 수 있도록 구성하였다.

$$\arg\max_{n}\left\{\omega\frac{V_{sei}}{V_{tot}}-(1-\omega)\frac{n}{N}\right\}. \tag{11}$$

위 식에서 V_{tot}와 V_{sel}은 각각 전체 데이터와 변수 선택 과정을 거친 데이터를 이용하여 얻은 공분산 행렬의 고유값 총합을 나타낸다. 이는 공분산 행렬의 고유값이 설명된 분산의 양을 의미하기 때문에 V_{sel}/V_{tot}는 전체 데이터가 갖는 분산 정보의 총량 대비 차원감소를 거친 데이터의 분산정보의 비율을 나타낸다. ω는 파라미터로서 분산정보와 선택된 변수의 수에 대한 가중치를 의미하는데 ω가 커질수록 분산정보에 대한 가중치가 더 커서 변수를 더 많이 선택하게 되며 반대로 작아질수록 선택한 변수에 대한 비용이 커져 적은 수의 변수를 선택하게 된다.

4. 시뮬레이션

데이터는 http://cs.joensuu.fi/sipu/datasets/에 공개된 군집화를 위한 시뮬레이션 데이터를 사용하였다. 데이터는 총 6개의 셋으로 구성되었으며, 모두 16개의 군집이 존재하며 1,024개의 샘플을 갖고 있다. 각 셋은 서로 다른 차원을 갖고 있다. 각 데이터 셋을 주성분분석을 이용하여 2개의 축으로 나타낸 스코어 그래프인<Figure 2>를 통해 16개의 군집을 명확하게 확인할 수 있다. 제안 기법을 검증하기 위한 지표로는 Representation entropy(RE)와 실루엣(Silhouette)을 이용하였다.

4.1 Representation Entropy(RE)

데이터가 p개의 변수를 갖고 있을 때, $p\times p$ 공분산 행렬의 고유값을 $\lambda_i(i, 1, 2, \cdots, p)$라 한다면, RE 지표는 다음과 같이 표현할 수 있다.

$$\mathrm{H_R}=-\sum_{i=1}^{p}\widetilde{\lambda}_i\log_k\widetilde{\lambda}_i \tag{12}$$

$$\text{whrer } \widetilde{\lambda}_i=\frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i}.$$

RE 지표는 공분산 행렬의 고유값을 이용하여 엔트로피를 계산한 것으로 고유값의 복잡도를 측정한 값이다. RE 지표에 대한 이해를 돕기 위해 이 값이 극단적으로 작은 경우를 생각한다면 공분산 행렬의 고유값이 하나를 제외한 나머지 모두 0을 갖는 경우에 가능하다. 즉 모든 변수가 선형관계가 강하기 때문에 하나의 선형 결합을 통해 표현 가능한 경우를 의미한다. 반면에 RE 지표의 값이 커짐에 따라 공분산 행렬의 고유값에 0이 아닌 값이 점점 늘어나며, 데이터를 재표현 하기 위해 여러 개의 변수 선형 결합이 필요하다는 것을 의미한다. 따라서, RE 지표 값이 클수록 차원 감소의 가능성이 낮으며, 값이 작을수록 차원감소의 가능성이 높음을 의미하며 동시에 데이터에 중복성(redundancy)이 높다고 해석할 수 있다(Mitra *et al*., 2002). 본 시뮬레이션에서는 동일한 데이터 셋에 대하여 선택한 변수의 수를 줄여 가며 지표를 관찰하였다. 따라서, 선택된 변수의 수가 많을 수록 RE 지표의 값이 커지는 경향이 있다. 이에 대한 보정을 위해 로그의 밑(k)을 선택한 변수의 수로 결정하여 RE 지표의 값을 0과 1사이의 값을 얻을 수 있도록 하였다.

4.2 실루엣(Silhouette)

실루엣은 군집화 결과에 대한 검증을 위한 지표로서 다음과

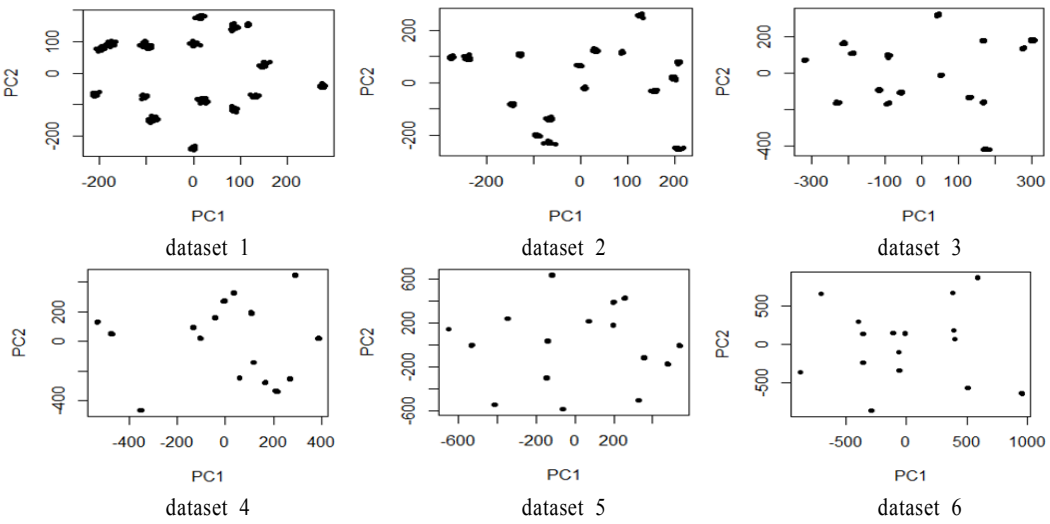

Figure 2. PCA score plots with two principal components for six simulation datasets

같이 표현할 수 있다.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (13)$$

위 식에서 $a(i)$는 i번째 샘플이 속한 군집내의 다른 샘플과 평균 거리를 의미한다. $b(i)$는 i번째 샘플이 속한 군집을 제외한 나머지 모든 샘플과의 평균 거리이다. $s(i)$는 -1과 1사이의 값을 가지며, 값이 클수록 군집화가 잘 되었음을 의미한다. 따라서, $s(i)$는 모든 샘플에 대해서 얻을 수 있으며, 이 값들의 평균을 통하여 군집화 결과를 검증할 수 있다(Bolshakova *et al*., 2003).

4.3 시뮬레이션 결과

수행한 시뮬레이션은 6개의 데이터에 대해 파라미터를 0.5와 0.3으로 바꿔가면서 변수 선택을 하지 않은 원래의 데이터셋과 RE 지표, 실루엣을 비교하였다.

결과를 요약하면 다음과 같다. 적은 값의 ω를 사용할 때 선택된 변수의 개수가 작아지는 것을 확인 할 수 있다. <Table 1>을 통해 제안 기법에 의해 축소된 차원의 RE를 보여주고 있으며 선택된 변수가 작아질수록 RE 지표 역시 높아지는 경향을 확인할 수 있다. 결과적으로 제안 변수선택 기법에 의해 중복성이 효과적으로 제거되었음을 보여주고 있다. ω가 0.3인 경우 1, 2번 데이터 셋의 결과에서 RE 지표가 결측치 인데, 이는 선택된 변수가 하나이기 때문에 고유값을 얻을 수 없기 때문이다.

제안 변수 선택기법이 군집화 문제에도 효과적으로 적용되는지 검증하기 위해 k-평균 알고리즘을 수행하여 얻은 결과를 실루엣 지표로써 관찰하였다. 군집화 기법으로 k-평균 군집화 알고리즘을 선택한 이유는 군집화를 위해 거리측도를 사용하기 때문이다. 차원이 커질 경우 거리 측도가 부정확 해지는 문제가 발생하게 되는데 제안 기법으로 차원을 감소함으로써 k-평균 군집화의 성능이 개선되는 것을 보이기 위해서이다. 단, k-평균 군집화 알고리즘의 경우 초기해에 따라 군집화의 결과가 달라지기 때문에 100회 반복 수행하여 군집을 결정하였다. <Table 1>의 마지막 열인 실루엣은 선택 변수의 수를 줄여가며 수행한 군집화 결과를 보여주고 있으며 대부분 제안 방법을 이용하여 변수 선택과정을 거친 데이터의 중복성이 제거 되어 더 좋은 군집화 결과를 확인할 수 있다.

5. 실제 데이터를 이용한 실험

5.1 데이터

제안 변수선택 기법의 효과를 검증하기 위해 실제 데이터를 이용한 실험에는 대표적인 고차원 데이터인 마이크로어레이(Microarray)데이터 14개를 이용하였다. 각 데이터셋은 DNA 정보를 담고 있는 변수와 질병의 발현 여부와 관련된 클래스 정보를 갖고 있어 분류문제에 적용할 수 있다. 사용한 데이터의 출처와 관련 질병은 <Table 2>에 실험결과와 함께 기록하였다.

5.2 실험 결과

실험은 제안 기법을 통해 변수를 선택했을 때, 데이터의 중복성이 효과적으로 제거되는지 RE 지표를 이용하여 평가하였고, 축소된 변수가 유의미한 변수인지 확인하기 위해 분류정확도를 살펴보았다. 분류 정확도는 k-인접 이웃 분류기를 100회 반복하여 분류 정확도의 평균과 표준편차를 표기하였다. 각 시행마다 트레이닝 셋은 50퍼센트를 무작위로 선택하여 구성하였으며 k-인접 이웃 분류기의 파라미터 k는 전체 관측치 수의 제곱근을 사용하였다(Mitra *et al*., 2002). k-인접 이웃 분류기를 사용한 이유는 시뮬레이션에서 군집화 알고리즘의 선택과 마찬가지로 차원이 커짐에 따라 부정확 해지는 거리측도의 성능이 제안 기법으로 차원을 줄였을 때 성능이 개선되는지 비교하기 위해서다.

실험결과는 <Table 2>에 기록하였으며, 비교는 앞서 시뮬레이션과 마찬가지로 모든 변수를 사용한 데이터셋과 제안 기법의 파라미터를 0.5와 0.3로 설정했을 때 축소된 데이터셋을 비교하였다. 파라미터를 0.5로 설정하였을 때 선택된 변수는 평균적으로 전체 변수의 29.75%로 줄었고, 파라미터를 0.3으로

Table 1. A simulation results for each dataset. It shows the representation entropy and silhouette of each dataset generated by proposed feature selection algorithm while changing parameter ω

Dataset	Dimension			Representation Entropy(RE)			Silhouette		
	Total	ω = 0.5	ω = 0.3	Total	ω = 0.5	ω = 0.3	Total	ω = 0.5	ω = 0.3
1	32	14	1	0.684	0.790	NA	0.591	0.514	0.597
2	64	25	1	0.594	0.699	NA	0.607	0.600	0.690
3	128	36	10	0.529	0.648	0.759	0.739	0.585	0.507
4	256	71	16	0.470	0.570	0.763	0.568	0.466	0.611
5	512	152	31	0.425	0.493	0.663	0.631	0.799	0.680
6	1,024	257	33	0.381	0.451	0.655	0.586	0.602	0.694

Table 2. An experiment results for each dataset. Author is reference information about published the paper using dataset. N is the number of observations and P is a dimension of each dataset. RE indicates representation entropy. KNNA is an accuracy of knn classifier, which include mean and standard deviation

Author	Disease	N	Parameter	P	RE	KNNA	
						Mean	SD
Alon *et al.*(1999)	Colon cancer	62	Total	2,000	0.81	73.06	8.64
			0.5	378	0.82	72.48	8.23
			0.3	182	0.80	69.84	7.95
Borovecki *et al.*(2005)	Huntington's disease	31	Total	22,283	0.62	59.93	12.41
			0.5	20,365	0.61	59.60	12.38
			0.3	5	0.77	40.00	11.13
Chin *et al.*(2006)	Breast cancer	118	Total	22,215	0.57	85.47	4.03
			0.5	4,399	0.64	83.54	5.14
			0.3	1,138	0.68	73.49	7.81
Chowdary *et al.*(2006)	Breast cancer	104	Total	22,283	0.71	85.33	5.04
			0.5	7,363	0.81	88.08	4.84
			0.3	576	0.82	69.44	6.53
Gordon *et al.*(2002)	Lung cancer	181	Total	12,533	0.83	89.22	3.75
			0.5	2,528	0.90	92.43	4.23
			0.3	176	0.84	84.55	4.11
Gravier *et al.*(2010)	Breast cancer	168	Total	2,905	0.93	66.54	4.01
			0.5	657	0.94	66.46	3.97
			0.3	189	0.91	66.36	4.05
Khan *et al.*(2001)	SRBCT	63	Total	2,308	0.89	38.42	6.76
			0.5	627	0.90	39.77	8.64
			0.3	159	0.88	35.77	7.18
Pomeroy *et al.*(2002)	CNS tumor	60	Total	7,128	0.86	58.83	7.92
			0.5	3,136	0.86	59.00	7.98
			0.3	293	0.80	51.07	10.39
Shipp *et al.*(2002)	Lymphoma	58	Total	7,129	0.80	86.67	7.66
			0.5	1,620	0.77	86.95	7.68
			0.3	197	0.74	73.05	6.93
Singh *et al.*(2002)	Prostate cancer	102	Total	12,600	0.79	77.59	6.09
			0.5	2,469	0.80	77.45	6.91
			0.3	368	0.83	72.33	8.38
Subramanian *et al.*(2005)	Cancer	50	Total	10,110	0.64	62.36	7.15
			0.5	2,623	0.69	59.64	9.41
			0.3	391	0.73	60.96	7.90
Tian *et al.*(2003)	Myeloma	173	Total	12,625	0.85	79.14	3.03
			0.5	829	0.87	78.48	3.33
			0.3	65	0.71	78.92	3.04
West *et al.*(2001)	Breast cancer	49	Total	7,129	0.88	51.64	8.62
			0.5	2,486	0.92	49.64	7.57
			0.3	329	0.92	45.96	7.99

설정했을 때는 평균적으로 3.88%로 줄었다. 이때 데이터의 중복성을 나타내는 RE 지표는 축소된 데이터가 대부분 모든 변수를 사용했을 때보다 증가하여 제안 기법에 의해 데이터의 중복성이 줄어들었음을 확인할 수 있었다. 단, Lymphoma 데이터의 경우에는 축소한 데이터의 중복성이 더 커지는 경우가 예외적으로 발생하였다. 또한, 제안 기법의 파라미터를 0.3으로 설정하였을 때 축소된 데이터의 중복성이 더 커지는 현상이 발생함을 알 수 있다. 이와 같은 현상은 공분산 행렬의 고유

값을 얻는 과정에서 데이터에 대한 선형변환이 일어나게 되는데, 이때 축소된 변수들간의 선형결합이 더 많은 변수들 사이의 선형결합보다 더 강하기 때문에 나타날 수 있는 현상이다. 이런 현상은 제안 하는 변수 선택기법이 주성분 분석을 바탕으로 변수를 선택하는 데, 주성분 분석은 데이터의 비선형 분포를 모델링 하는데 적합하지 않기 때문에 발생할 수 있다.

제안기법을 이용하여 축소한 데이터의 분류 성능은 전체적으로 더 우수하거나 모든 변수를 사용했을 때 와 비슷한 수준을 보이는 것을 확인할 수 있었다. 이는, 제안 기법이 클래스를 분류하는 중요한 변수를 올바르게 선택했음을 의미한다. 본 연구는 예측 모델 기반의 지도 변수 기법이 아닌 비지도 변수 선택기법임을 고려한다면, 데이터의 중복성을 줄이는 동시에 분산 정보 손실을 최소화 할 수 있는 변수를 선택함으로써 얻는 부가적인 결과로 해석할 수 있다. 그리고 축소된 데이터 간의 분류 성능을 비교하면, 대부분의 경우 파라미터를 0.5로 설정한 경우에 더 좋은 성능을 나타냈다. 파라미터를 0.3으로 설정하여 더 적은 수의 변수를 사용하였을 때는 상대적으로 분류 성능이 저하되는 결과를 얻었다.

6. 결 론

최근 계측기술과 데이터 저장기술이 발달함에 따라 대용량의 고차원 데이터가 곳곳에서 쏟아져 나오고 있다. 이와 같은 대용량 고차원 데이터를 효과적으로 분석하기 위해서는 차원축소가 필수적이다. 본 논문에서는 차원 축소 기법 분야 중 상대적으로 연구가 덜 행해진 비지도 변수 선택 기법에 대해 살펴보았으며 새로운 기법을 제안하였다.

제안기법은 주성분 분석으로부터 얻은 로딩벡터의 기하학적 정의를 근간으로 하고 있다. 즉, 모든 변수가 동일한 중요도를 갖는 벡터를 정의하고 실제 데이터로부터 구한 로딩벡터와의 비교를 통해 중요 변수를 찾을 수 있었다.

제안기법은 시뮬레이션 데이터와 실제 데이터를 이용해 RE 지표를 통해 데이터의 중복성을 줄일 수 있음을 입증하였고, 대표적인 비지도 학습인 군집화 문제에 효과적으로 적용 가능함을 실루엣을 통해 보였다. 또한, 마이크로어레이 데이터를 이용한 실제 데이터에 대해 수행한 분류 성능 검증을 통해 효용성과 실제문제 적용성을 확인할 수 있었다. 하지만, 주성분 분석을 바탕으로 유도한 제안 기법은 비선형 분포를 갖는 데이터에 대해 성능에 문제를 갖고 있다. 좀 더 구체적으로 설명하자면, 주성분 분석은 비선형 분포를 갖는 데이터에 효과적으로 적용되지 않기 때문에 로딩벡터의 신뢰도가 떨어지게 되고, 이를 바탕으로 변수를 선택했을 때 오히려 데이터의 중복성이 커지는 결과를 얻을 수 있다. 이러한 문제점은 주성분 분석을 사용하는데 따른 효과이며 동시에 데이터의 구조를 선형으로 근사하여 모델링 하여 발생하는 제안기법의 한계점이다.

추후 연구방향으로는 비선형 데이터에 대해서 커널을 이용한 주성분 분석을 적용하는 것과 함께 제안 기법의 최적 파라미터 선정 문제를 남겨놓고 있다. 우선, 비선형 데이터에 대해 발생하는 문제는 커널(kernel) 등의 방법을 사용하는 비선형 주성분 분석을 적용해 볼 수 있다. 또한, 앞서 언급했듯이 제안기법은 파라미터에 따라 데이터의 중복성과 분류 성능이 차이가 있음을 알 수 있다. 또한, 축소된 차원도 파라미터의 변화와 선형관계에 있지 않음을 알 수 있다. 즉, 파라미터에 따라 선택되는 변수의 수가 큰 차이를 나타냄을 의미한다. 이는 변수의 중요도가 선형관계를 갖지 않기 때문이므로, 변수 선택의 기준이 되는 식 (8)의 최적해를 구하기 위한 방법을 중요도의 분포를 반영하여 더 효과적으로 수행할 수 있도록 개선할 여지가 있다. 그리고 식 (8)의 개선에 최적의 파라미터를 결정하는 문제를 추가하여 파라미터의 변동에도 비교적 강건한 (robust) 변수 선택 기법에 대한 연구를 남겨놓고 있다.

참고문헌

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999), Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proceedings of the National Academy of Sciences*, **96**(12), 6745-6750.

Bolshakova, N. and Azuaje, F. (2003), Cluster validation techniques for genome expression data, *Signal processing*, **83**(4), 825-833.

Borovecki, F., Lovrecic, L., Zhou, J., Jeong, H., Then, F., Rosas, H. D., and Krainc, D. (2005), Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease, *Proceedings of the National Academy of Sciences of the United States of America*, **102**(31), 11023-11028.

Boutsidis, C., Mahoney, M. W., and Drineas, P. (2008), Unsupervised feature selection for principal components analysis, *In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, *ACM*, 61-69.

Dash, M., Choi, K., Scheuermann, P., and Liu, H. (2002), Feature selection for clustering-a filter solution. *In Data Mining*, *2002*, *ICDM 2003*, *Proceedings*, *2002 IEEE International Conference*, *IEEE*, 115-122

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W. L., and Gray, J. W. (2006), Genomic and transcriptional aberrations linked to breast cancer pathophysiologies, *Cancer cell*, **10**(6), 529-541.

Chowdary, D., Lathrop, J., Skelton, J., Curtin, K., Briggs, T., Zhang, Y., and Mazumder, A. (2006), Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative, *The journal of molecular diagnostics*, **8**(1), 31-39.

Gordon, G. J., Jensen, R. V., Hsiao, L. L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., and Bueno, R. (2002), Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma, *Cancer research*, **62**(17), 4963-4967.

Gravier, E., Pierron, G., Vincent-Salomon, A., Gruel, N., Raynal, V., Savignoni, A., and Delattre, O. (2010), A prognostic DNA signature for T1T2 node-negative breast cancer patients, *Genes, Chromosomes*

and Cancer, **49**(12), 1125-1134.

Guo, Q., Wu, W., Massart, D. L., Boucon, C., and De Jong, S. (2002), Feature selection in principal component analysis of analytical data, *Chemometrics and Intelligent Laboratory Systems*, **61**(1), 123-132.

Guyon, I. and Elisseeff, A. (2003), An introduction to variable and feature selection, *The Journal of Machine Learning Research*, **3**, 1157-1182.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009), *The elements of statistical learning*, **2**(1), New York : Springer.

Jolliffe, I. T. (1972), Discarding variables in a principal component analysis, I : Artificial data. *Applied statistics*, 160-173.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., and Meltzer, P. S. (2001), Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature medicine*, **7**(6), 673-679.

Kim, S. B. (2009), Feature Extraction/Selection in High-Dimensional Spectral Data, In J. Wang (Ed.), *Encyclopedia of Data Warehousing and Mining*, *Second Edition*, (863-869), Hershey, PA : Information Science Reference, doi:10.4018/978-1-60566-010-3.ch133.

Kim, S. B. and Rattakorn, P. (2011), Unsupervised feature selection using weighted principal components, *Expert Systems with Applications*, **38**(5), 5704-5710.

Malhi, A. and Gao, R. X. (2004), PCA-based feature selection scheme for machine defect classification, *Instrumentation and Measurement*, *IEEE Transactions*, **53**(6), 1517-1525.

Mao, K. Z. (2005), Identifying critical variables of principal components for unsupervised feature selection, *Systems, Man, and Cybernetics, Part B : Cybernetics*, *IEEE Transactions*, **35**(2), 339-344.

Mitra, P., Murthy, C. A., and Pal, S. K. (2002), Unsupervised feature selection using feature similarity, *IEEE transactions on pattern analysis and machine intelligence*, **24**(3), 301-312.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., and Golub, T. R. (2002), Prediction of central nervous system embryonal tumour outcome based on gene expression, *Nature*, **415**(6870), 436-442.

Roth, V. and Lange, T. (2003), Feature selection in clustering problems, *In Advances in neural information processing systems*.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., and Golub, T. R. (2002), Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, *Nature medicine*, **8**(1), 68-74.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., and Mesirov, J. P. (2005), Gene set enrichment analysis : a knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the National Academy of Sciences of the United States of America*, **102**(43), 15545-15550.

Tian, E., Zhan, F., Walker, R., Rasmussen, E., Ma, Y., Barlogie, B., and Shaughnessy Jr, J. D. (2003), The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma, *New England Journal of Medicine*, **349**(26), 2483-2494.

Wang, P. and Kim, J. (2014), Analysis of Chinese Provinces for Introduction of Reverse Mortgage Scheme Using Principal Component Analysis, *Journal of the Korean Institute of Industrial Engineers*, **40**(2), 205-214.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., and Nevins, J. R. (2001), Predicting the clinical status of human breast cancer by using gene expression profiles, *Proceedings of the National Academy of Sciences*, **98**(20), 11462-11467.

Widjaja, D., Varon, C., Dorado, A., Suykens, J. A., and Van Huffel, S. (2012), Application of Kernel Principal Component Analysis for Single-Lead-ECG-Derived Respiration, Biomedical Engineering, *IEEE Transactions on*, **59**(4), 1169-1176.

Yu, L. and Liu, H. (2003), Feature selection for high-dimensional data : A fast correlation-based filter solution, *In ICML*, **3**, 856-863.