The Korean Journal of Applied Statistics
2023, Vol. 36, No. 6, 501–514

DOI: https://doi.org/10.5351/KJAS.2023.36.6.501
Print ISSN1225-066X / Online ISSN 2383-5818

Feature selection for text data via sparse principal component analysis

Won Son[1,a]

[a]Department of Information Statistics, Dankook University

Abstract

When analyzing high dimensional data such as text data, if we input all the variables as explanatory variables, statistical learning procedures may suffer from over-fitting problems. Furthermore, computational efficiency can deteriorate with a large number of variables. Dimensionality reduction techniques such as feature selection or feature extraction are useful for dealing with these problems. The sparse principal component analysis (SPCA) is one of the regularized least squares methods which employs an elastic net-type objective function. The SPCA can be used to remove insignificant principal components and identify important variables from noisy observations. In this study, we propose a dimension reduction procedure for text data based on the SPCA. Applying the proposed procedure to real data, we find that the reduced feature set maintains sufficient information in text data while the size of the feature set is reduced by removing redundant variables. As a result, the proposed procedure can improve classification accuracy and computational efficiency, especially for some classifiers such as the k-nearest neighbors algorithm.

Keywords: elastic net, feature selection, regularization method, sparse principal component analysis

1. 서론

변수의 개수가 많은 데이터를 분석할 때 모든 변수를 분석에 활용하는 경우 과적합 문제가 발생하기 쉽다. 특히 k-최근접이웃(k-nearest neighbors; KNN) 알고리즘 등 일부 분류모형의 경우 변수가 많아질수록 잡음이 누적되어 분류 성능이 저하되는 현상을 확인할 수 있다. 따라서 변수가 지나치게 많은 경우에는 불필요하거나 중요도가 낮은 변수를 제거하여 변수들을 소수의 새로운 변수로 요약하는 차원축소(dimensionality reduction) 과정을 거치기도 한다. 이렇게 일부 변수를 제거하는 과정을 변수선택 또는 특징선택(feature selection)이라 하고, 변수들을 합성하여 소수의 새로운 변수들로 데이터를 요약하는 과정을 특징추출(feature extraction)이라 부르기도 한다.

주성분분석(principal component analysis; PCA)은 대표적인 특징추출 방법 중 하나로 회전변환을 통해 데이터의 변동을 가장 잘 설명할 수 있는 성분들을 식별하는 과정이다. 주성분분석에서 데이터의 변동을 가장 잘 설명하는 성분들은 데이터행렬 $\mathbf{X}$의 특잇값 분해(singular value decomposition)에서 큰 특잇값(singular value)들에 대응되는 특이벡터(singular vector)들에 해당된다. 즉, 관측값의 개수가 n, 변수의 수가 p인 $n \times p$ 차원 데이터행렬 $\mathbf{X}$가 $n \times n$ 차원 직교행렬 $\mathbf{U}$와 $p \times p$ 차원 직교행렬 $\mathbf{V}$, $n \times p$ 차원 직사각대각행렬(rectangular diagonal matrix) $\mathbf{D}$에 대해 $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\intercal}$로 분해될 때 주성분분석을 통해 데이터의 차원을 축소하기 위해서는

[1]Department of Information Statistics, Dankook University, 152 Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do 16890, Korea. E-mail: son.won@dankook.ac.kr

Published 31 December 2023 / journal homepage: http://kjas.or.kr

행렬 $\mathbf{D}$의 대각원소 $d_1, d_2, \ldots, d_p$ 중 값이 큰 q개의 대각원소 $d_{(n)}, d_{(n-1)}, \ldots, d_{(n-q+1)}$를 선택하여 데이터를 요약한다. 이 때 이들 대각원소에 대응되는 행렬 $\mathbf{V}$의 열들이 데이터를 요약하는 q개의 축을 생성하는 회전변환에 해당된다.

Zou 등 (2006)은 주성분의 수와 각 주성분의 선형식에서 0이 아닌 계수에 희소성을 부여하기 위한 방법으로 희소주성분분석(sparse principal component analysis; SPCA)을 제안한 바 있다. 주성분분석을 통해 많은 변수로 이루어진 고차원 데이터의 차원을 축소할 수 있지만 주성분분석에서 식별된 각각의 주성분들은 여전히 모든 변수들의 선형결합으로 이루어져 있어 각 주성분의 해석에도 주관성이 개입될 수 밖에 없다는 문제점이 있다. 즉, 중요도가 낮은 변수들이 제거되지 않고 여전히 주성분분석으로 생성된 각 주성분에 포함되어 있으며 이 변수들이 주성분이 어떤 의미를 가지는지 해석하기 어렵게 하는 요인이 될 수 있다. SPCA에서는 제약이 부여된(regularized) 최소제곱법을 통해 중요도가 낮은 주성분을 제거하고 각 주성분에서도 중요도가 낮은 변수들을 제거하여 주성분의 수와 주성분에 사용되는 변수의 수를 축소한다.

텍스트데이터는 일반적으로 많은 단어들로 구성되므로 각 단어를 하나의 변수로 간주할 때 변수가 많은 데이터에 해당된다. 텍스트데이터의 경우에도 주성분분석과 같은 특징추출 방법을 사용할 수 있지만 각 단어들이 고유한 의미를 내포하고 있다는 점에서 각 단어의 원형을 보존한 상태로 데이터를 분석하는 것을 선호하는 경우도 있다. 이러한 이유 때문에 텍스트데이터에서 중요도가 낮은 단어들을 제거하는 다양한 절차들이 제안된 바 있다. 자주 사용되는 단어선택 절차는 Yang과 Pedersen (1997), Mladenić와 Grobelnik (1999), Forman (2003), Chen 등 (2009)에서 확인할 수 있다. 이 방법들 중 단어출현빈도와 범주 사이의 관계를 요약한 이차원분할표를 기반으로 하는 방법들이 자주 사용된다. 한편, 이 방법들은 주로 텍스트데이터의 목표변수로 사용되는 범주에 대한 정보를 활용하는 방법들이므로 범주와 연관된 정보를 가진 단어들만 선택될 가능성이 있다. 이러한 점을 고려하여 Jang 등 (2022)은 텍스트데이터에 내재된 토픽을 식별하는 데 사용되는 잠재디리클레할당(latent Dirichlet allocation; LDA)을 활용하여 범주 정보에 의존하지 않는 단어선택 방법을 제안하였다.

이 연구에서는 범주에 대한 정보를 사용하지 않고 SPCA를 이용하여 텍스트데이터의 분석에 유용한 단어들을 선택하는 방법을 제안한다. SPCA에서 선택된 주성분들은 텍스트데이터를 잘 요약하는 정보를 가지고 있으며 이 주성분들을 구성하는 변수들, 즉 단어들 또한 텍스트데이터를 잘 대표하는 단어들이라고 볼 수 있다. 이 연구는 다음과 같이 구성된다. 2절에서는 SPCA와 관련된 선행연구들을 정리해보고 SPCA를 이용한 단어선택 절차를 제안한다. 3절에서는 실제 데이터에 SPCA를 적용하여 단어를 선택하고 선택된 단어들을 이용하여 분류분석을 실시해본다. 분류분석 결과를 통해 범주에 대한 정보를 활용하지 않고 제안된 절차를 통해 선택된 단어들이 분류분석에 충분한 정보를 보존하고 있는지 확인해본다. 마지막으로 4절에서는 결과를 정리하고 앞으로의 연구과제를 소개한다.

2. 희소주성분분석을 이용한 텍스트데이터의 단어 선택

2.1. 희소주성분분석

1절에서 간략히 살펴본 바와 같이 주성분은 데이터행렬 $\mathbf{X}$의 특잇값분해 $\mathbf{X} = \mathbf{UDV}^\intercal$를 통해 구할 수 있다. 식 $\mathbf{X} = \mathbf{UDV}^\intercal$에서 $\mathbf{XV} = \mathbf{UD}$이므로 데이터행렬 $\mathbf{X}$를 직교행렬 $\mathbf{V}$로 회전변환한 것이 행렬 $\mathbf{UD}$에 해당된다. 즉, 데이터행렬 $\mathbf{X} = [\mathbf{x}_1^\intercal, \mathbf{x}_2^\intercal, \ldots, \mathbf{x}_n^\intercal]^\intercal$의 i 번째 행에 해당되는 p차원 관측값 $\mathbf{x}_i$는 직교행렬 $\mathbf{V} = [V_1, V_2, \ldots, V_p]$에 의해 p개의 축을 가진 새로운 직교좌표공간으로 변환된다. 여기서 $\mathbf{x}_i$ $(i = 1, 2, \ldots, n)$는 데이터행렬 $\mathbf{X}$의 행벡터, V_j $(j = 1, 2, \ldots, p)$는 회전변환 행렬 $\mathbf{V}$의 열 벡터를 의미한다.

데이터 $\mathbf{X}$를 q개의 변수로 요약하기 위해서는 직사각대각행렬 $\mathbf{D}$에서 값이 큰 q개의 대각원소에 대응되는 행렬 $\mathbf{V}$의 q개의 열 벡터를 선택하여 작성한 $p \times q$ 행렬 $\mathbf{V}_q$를 이용할 수 있다. 즉, q개의 주성분으로 데이터를 요약하는 과정은 행렬 $\mathbf{X}$와 $\mathbf{XV}_q\mathbf{V}_q^\intercal$의 차이가 최소가 되도록 하는 행렬 $\mathbf{V}_q$를 찾는 과정으로 볼 수 있으며 이

Table 1: Example of the principal component analysis

	Original data						Noisy observations								
	(a) Data			(b) Rotated data			(c) Data			(d) Rotated data			(e) Sparse rotated data		
	x_1	x_2	x_3	PC1	PC2	PC3	x_1	x_2	x_3	PC1	PC2	PC3	SPC1	SPC2	SPC3
obs1	2	1	0	2.236	0.000	0.000	1.374	1.330	0.576	1.820	0.818	0.066	1.766	0.000	0.000
obs2	4	2	0	4.472	0.000	0.000	4.184	1.180	−0.305	4.281	−0.775	−0.249	4.325	0.000	0.000
obs3	6	3	0	6.708	0.000	0.000	5.164	3.487	1.512	6.185	1.666	−0.288	6.084	0.000	0.000
obs4	8	4	0	8.944	0.000	0.000	9.595	4.738	0.390	10.705	0.272	0.004	10.665	0.000	0.000

과정은

$$\widehat{\mathbf{V}}_q = \arg\min_{\mathbf{A}} \sum_{i}^{n} \|\mathbf{x}_i - \mathbf{x}_i\mathbf{A}\mathbf{A}^\intercal\|_2^2 \tag{2.1}$$

과 같은 제곱합의 최소화 문제로 표현할 수 있다. 식 (2.1)에서 행렬 $\mathbf{A}$는 각 열들이 서로 수직인 $n \times q$ $(q \le p)$ 차원 직교행렬이고 $\|\mathbf{a}\|_2$는 벡터 $\mathbf{a}$의 ℓ_2-노름을 의미한다.

Zou 등 (2006)은 $\mathbf{B} = [\beta_1, \beta_2, \ldots, \beta_p]$라 할 때 β_j $(j = 1, 2, \ldots, p)$에 대한 ℓ_2-노름 제곱의 λ배를 오차제곱합에 더한 식

$$\left(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}\right) = \arg\min_{\mathbf{A},\mathbf{B}} \left\{ \sum_{i}^{n} \|\mathbf{x}_i - \mathbf{x}_i\mathbf{B}\mathbf{A}^\intercal\|_2^2 + \lambda \sum_{j=1}^{q} \left\|\beta_j\right\|_2^2 \right\} \tag{2.2}$$
$$\text{subject to } \mathbf{A}^\intercal\mathbf{A} = \mathbf{I}_q$$

를 최적화하여 해 $\widehat{\mathbf{A}}, \widehat{\mathbf{B}}$을 구하면 $\widehat{\mathbf{B}} = [\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_q]$의 각 열이 PCA에서 구해지는 회전변환 행렬 $\mathbf{V}_q = [V_1, V_2, \ldots, V_q]$에 대해 $\hat{\beta}_j \propto V_j$임을, 즉 $\mathbf{V}_q$의 각 열들의 상수 배로 표현될 수 있음을 보였다. 식 (2.2)의 손실함수는 능형회귀(ridge regression)와 같은 형태이며 식 (2.2)를 엘라스틱넷(elastic net)과 같은 형태로 변형하여

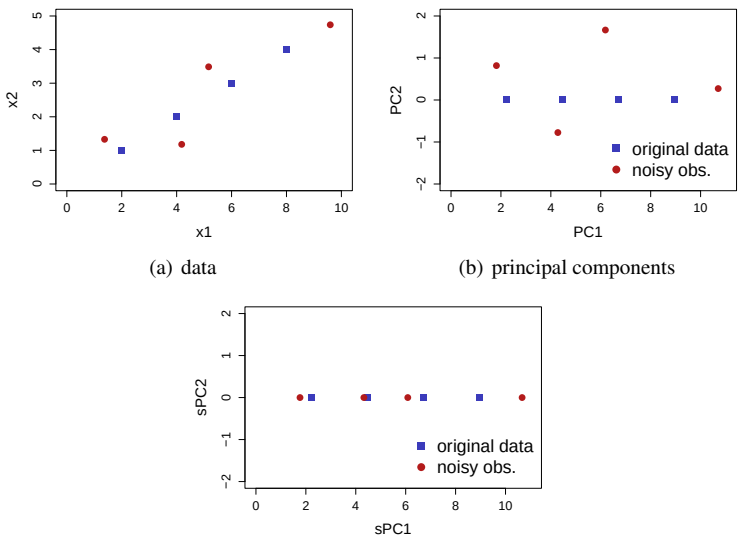
$$\left(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}\right) = \arg\min_{\mathbf{A},\mathbf{B}} \left\{ \sum_{i}^{n} \|\mathbf{x}_i - \mathbf{x}_i\mathbf{B}\mathbf{A}^\intercal\|_2^2 + \lambda \sum_{j=1}^{q} \left\|\beta_j\right\|_2^2 + \sum_{j=1}^{q} \lambda_{1,j} \left\|\beta_j\right\|_1 \right\} \tag{2.3}$$
$$\text{subject to } \mathbf{A}^\intercal\mathbf{A} = \mathbf{I}_q$$

와 같이 손실함수에 β_j $(j = 1, 2, \ldots, p)$에 대한 ℓ_1-노름을 추가하면 β_j의 희소성을 구현할 수 있게 된다 (Zou 등, 2006). 따라서 식 (2.3)을 최적화하여 구한 해 $\widehat{\mathbf{B}}$의 각 열은 많은 원소가 0이고 일부 원소만 0이 아닌 벡터로 표현될 수 있다.

일반적인 조건 아래에서 식 (2.3)의 해를 대수적인 방법으로 구하기는 어려우므로 수치적인 방법으로 해를 구한다. Zou 등 (2006)은 $\mathbf{B}$가 주어진 상태에서 $\mathbf{A}$의 최솟값을 구하고, 다시 $\mathbf{A}$가 주어진 상태에서 $\mathbf{B}$를 구하는 과정을 반복함으로써 식 (2.3)을 최적화하는 방법을 제안하였다. $\mathbf{B}$가 고정된 값으로 주어진 상태에서는 식 (2.3)의 벌점항을 무시할 수 있으므로 일반적인 주성분분석 과정으로 볼 수 있고, $\mathbf{A}$가 고정된 값으로 주어진 상태에서는 식 (2.3)은 $\mathbf{B}$에 관한 엘라스틱넷 해로 볼 수 있다. 이렇게 반복적인 수치해를 구하는 과정에서 $\mathbf{X}^\intercal\mathbf{X}\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^\intercal$로 분해할 수 있으며 행렬 $\mathbf{D}$를 이용하여 특잇값을 구할 수 있다.

2.2. 희소주성분분석의 예

예를 들어 Table 1의 (a)와 같이 세 변수 x_1, x_2, x_3와 네 개의 관측값으로 이루어진 데이터를 생각해보자. 이 데이터에서 잡음이 없는 원 데이터의 변수 x_1과 x_2는 $2x_2 = x_1$이라는 선형관계를 가지고 있고 변수 x_3의 값은



(a) data (b) principal components

(c) sparse principal components

Figure 1: *Example of observations, PCs and sparse PCs.*

Table 2: Examples of PC loadings for the original PCA and the sparse PCA

	(a) PC loadings for original data			(b) PC loadings for noisy data			(c) SPC loadings for noisy data		
	PC1	PC2	PC3	PC1	PC2	PC3	PC1	PC2	PC3
x_1	0.894	0.000	−0.447	0.904	−0.308	−0.295	0.930	0.000	0.000
x_2	0.447	0.000	0.894	0.427	0.623	0.656	0.367	0.000	0.000
x_3	0.000	1.000	0.000	0.018	0.719	−0.695	0.000	0.000	0.000

항상 0이다. 따라서 이 데이터를 회전변환하면 Table 1의 (b)에서와 같이 첫 번째 주성분만 0이 아닌 값을 가지고 다른 두 주성분은 항상 0인 값을 가진 형태로 변환됨을 알 수 있다. 다음으로 Figure 1(a)와 같이 잡음이 없는 관측값 주변에서 관찰된 잡음이 추가된 데이터를 생각해보자. 잡음이 추가된 데이터를 주성분분석을 통해 회전변환하여 표현하면 Table 1의 (d)와 Figure 1의 (b)에서와 같이 첫 번째 주성분 외의 다른 주성분들도 0이 아닌 값을 가지게 된다. 즉, 관측값에 오차가 추가되어 있는 경우 오차항으로 인해 0이 아닌 주성분 수가 많아진다.

Table 2의 (a)에서 잡음이 없는 데이터의 경우 각 주성분을 표현하는 선형모형의 계수 중 일부만 0이 아닌 값을 가지고 있음을 확인할 수 있다. 반면, Table 2의 (b)와 같이 잡음이 추가된 데이터에서는 각 주성분을 표현하는 선형관계식의 모든 계수들이 0이 아닌 값들로 이루어져 있음을 확인할 수 있다. 예를 들어 첫 번째 주성분의 경우 잡음이 없는 데이터에서는 변수 x_3의 계수가 0이지만 잡음이 추가되면 계수 값이 0.018로 변하는 것을 볼 수 있다. 또, 상대적으로 특잇값이 작은 두 번째와 세 번째 주성분의 경우 선형식의 계수가 더 큰 폭으로 변하는 것을 알 수 있다.

이렇게 잡음이 포함된 데이터에 주성분분석을 적용하는 경우 잡음이 없는 경우에 비해 각 주성분이 더 복잡한 선형관계식으로 표현되는데 희소주성분분석은 모형을 단순하게 표현할 수 있도록 해준다. 희소주성분분석은 R의 `elasticnet`패키지를 이용하여 계산할 수 있다. Table 1의 (e)에 제시된 희소주성분분석 결과는 잡음이 포함된 데이터 (c)를 `elasticnet`패키지의 `spca` 함수에 입력하여 계산한 희소주성분분석 결과이다.

이 결과는 spca 함수에서 주성분 수를 조절하는 인수 K의 값을 1로, 각 주성분에서 0이 아닌 선형계수의 개수를 조절하는 인수 para의 값을 0.5로 지정하여 계산하였다. 이 결과에서 특잇값이 0에 해당되는 두 번째와 세 번째 주성분의 선형식 계수들이 모두 0으로 추정되는 것을 볼 수 있다. 또, 0이 아닌 특잇값에 대응되는 선형관계식의 계수 중 변수 x_3에 해당되는 값이 0으로 변화하여 주성분의 선형관계식도 단순하게 표현됨을 확인할 수 있다.

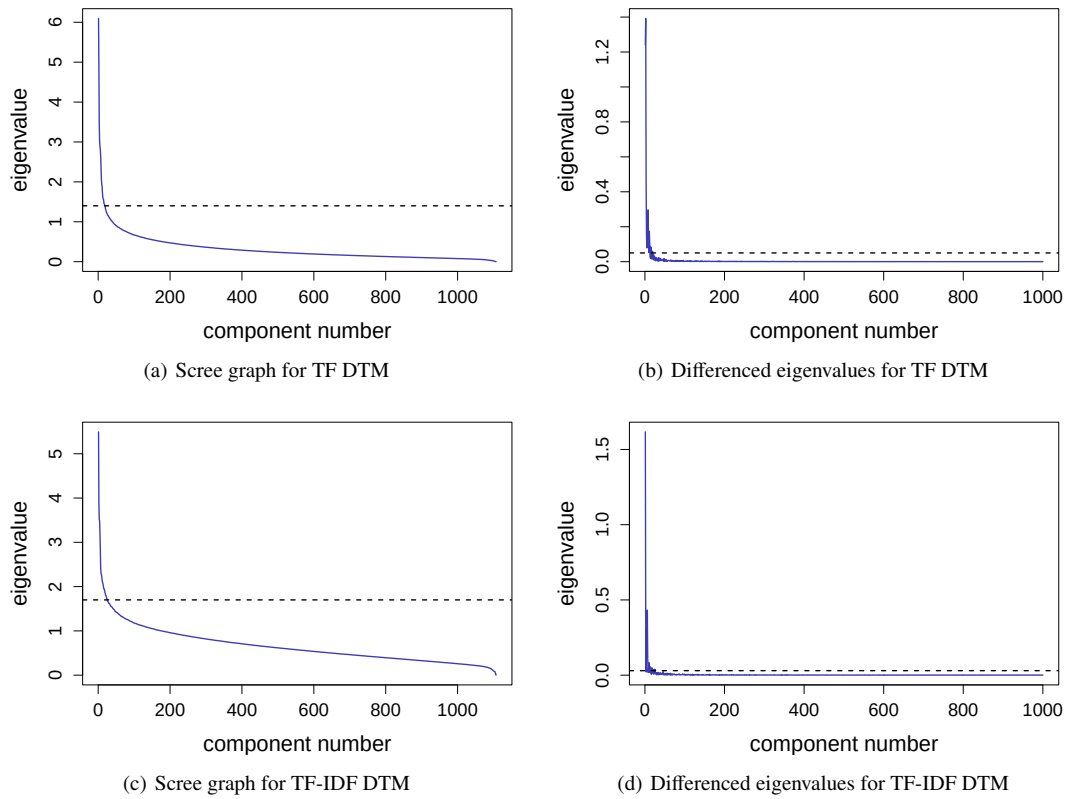
2.3. 희소주성분분석을 이용한 텍스트데이터의 단어선택 절차

지금까지 살펴본 바와 같이 희소주성분분석에서는 중요도가 높은 주성분을 제외한 나머지 주성분들은 제거되고 각 주성분을 표현하는 선형관계식의 계수들 중에서도 의미 있는 값들만 보존되는 특징이 있다. 즉 희소주성분분석을 이용하면 잡음에 해당되는 주성분을 제거하고 각 주성분에서도 유용하지 않은 변수를 제거할 수 있다. 이 절에서는 희소주성분분석의 이러한 특징을 이용하여 텍스트데이터에 포함된 많은 단어들 중에서 유용한 단어들만 선택하고 유용하지 않은 단어들은 제거하는 절차를 제안하려 한다.

희소주성분분석에서는 식 (2.3)의 조절모수 $\lambda_{1,j}$에 의해 주성분의 수와 각 주성분의 선형식에서 0이 아닌 계수를 가지는 변수들의 수가 결정된다. 한편, 적절한 주성분의 수와 계수가 0이 아닌 변수의 수를 결정하기 위해서 조절모수를 선택하려면 많은 반복 연산과정이 필요하다. elasticnet 패키지에서는 이러한 반복 연산과정을 거치지 않고도 주성분의 수 q와 함께 각 주성분에서 0이 아닌 변수의 수를 직접 지정할 수 있다. 따라서 이 연구에서는 먼저 일반적인 주성분분석을 통해 희소주성분분석에 적용할 적절한 주성분의 수 q를 정하고 elasticnet 패키지의 spca 함수에 주성분 수 K = q와 각 주성분의 선형관계식에서 0이 아닌 변수의 수를 입력하는 방식으로 단어를 선택하는 절차를 진행하였다. spca 함수의 para 인자에 벡터 값을 입력하여 각 주성분별로 벌점 또는 선택되는 변수의 수를 서로 다르게 지정할 수도 있지만, 이 연구에서와 같이 변수의 수가 많은 상황에서는 모든 경우의 수를 고려하기는 현실적으로 어려우므로 계산 상의 편의를 위해 각 주성분에서 동일한 수의 단어를 선택하였다.

제안된 절차에서 주성분의 수는 일반적인 주성분분석에서와 마찬가지로 정할 수 있다. 주성분의 수를 구하기 위한 방법으로는 총변동(total variation)의 일정 비율 이상을 설명하는 최소 개수의 주성분을 선택하는 방법, 스크리 그래프(scree graph)를 이용하는 방법 등이 있다 (Jolliffe, 2002). 이 연구에서는 2.1에서와 같은 방식으로 구한 특잇값에 대해 스크리 그래프를 적용하여 주성분 수를 결정하는 방법을 택하였다. 스크리 그래프는 i 번째로 큰 특잇값 d_i의 좌표를 (i, d_i)로 정의하고 이 좌표들을 순차적으로 연결한 그림으로 q 번째 특잇값 이전까지는 기울기가 가파르고 q 번째 특잇값 이후부터는 기울기가 완만해질 때 q를 주성분의 수로 결정하는 방법이다. 즉, 스크리 그림을 이용하여 주성분의 수를 구할 때 주성분의 수 q는 $i \le q$에 대해 $d_{i-1} - d_i$의 값과 $i \ge q$에 대해 $d_i - d_{i+1}$의 값이 상대적으로 큰 차이를 보이는 q를 주성분의 수로 선택한다.

한편, 주어진 주성분 수 q에 대해 각 주성분의 선형식에서 0이 아닌 계수를 가지는 변수의 수 m이 증가함에 따라 점차 많은 단어들이 선택되는데 이 단어들 중 일부 단어들은 여러 주성분에 공통적으로 포함되어 있을 수 있다. 이렇게 여러 주성분에 공통적으로 포함된 단어들이 늘어나면 각 주성분에서 선택되는 단어집합의 차이가 점차 줄어드는 것으로 해석할 수 있다. 즉, m이 증가함에 따라 선택되는 단어 수의 증가폭은 둔화되는데 이렇게 단어 수의 증가폭이 둔화되는 지점에 해당되는 m을 각 주성분 선형식의 계수가 0이 아닌 단어 수로 결정하였다. 즉, 주성분 q개에서 각각 m개의 단어를 선택하여 모두 qm개의 단어들이 선택되었을 때 이 중에서 u_m개의 단어가 중복되지 않은 단어라면 중복되지 않은 단어 비중은 $r_m = u_m/qm$으로 표현할 수 있다. $r_m - r_{m-1}$과 $r_{m+1} - r_m$의 값을 비교하여 두 값의 차이가 상대적으로 작아질 때의 m을 각 주성분에서의 적절한 단어 수로 결정하였다. 이렇게 중복된 단어 비중이 크게 증가하기 직전의 m 값으로 각 주성분에서 선택되는 단어 수를 결정한 것은 중복되는 단어 비중이 커질수록 추가되는 단어가 주는 정보가 많지 않을 것으로 판단하였기 때문이다. 이상의 과정은 다음과 같은 절차로 요약할 수 있다.



(a) Scree graph for TF DTM

(b) Differenced eigenvalues for TF DTM

(c) Scree graph for TF-IDF DTM

(d) Differenced eigenvalues for TF-IDF DTM

Figure 2: *Scree graphs.*

1. 텍스트데이터의 DTM에 PCA를 적용한다.
2. 스크리 그림을 이용하여 적절한 주성분의 수 q를 결정한다.
3. 주성분 수가 q로 주어졌을 때 각 주성분에서 0이 아닌 계수를 가지는 단어 수를 m ($m = 1, 2, \ldots$)으로 하여 희소주성분분석을 적용한다.
4. 3의 결과 중 여러 주성분에서 중복되어 선택되는 단어 비중이 크게 높아지기 직전의 m을 각 주성분에서 선택할 적절한 단어 수로 정한다.

3. 실제 데이터에의 적용

3.1. 분석 대상 데이터

분석 대상 실제 데이터로는 로이터(Reuters)-21578 데이터를 사용하였다. 로이터-21578 데이터는 로이터 통신사의 1987년 기사 데이터로 텍스트데이터 연구의 벤치마크 데이터로 널리 활용되고 있다 (Manning과 Schutze, 1999). 로이터-21578 데이터에는 총 10,788건의 기사가 포함되어 있는데 이 연구에서는 로이터-21578 데이터 중 기업인수(acq), 수익(earn), 무역(trade), 원유(crude), 외환(money-fx) 등 5개의 주제에 해당되는 기사 3,148건을 분석 대상 데이터로 선정하였다.

Table 3: Selected words for term frequency DTM (20 words for each principal component)

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
bank	**also**	bank	analyst	bank	bank	analyst	billion
billion	analyst	billion	bank	billion	company	bank	**deficit**
bpd	billion	company	billion	company	currency	**barrel**	export
company	**can**	**country**	company	cts	**dealer**	billion	fall
cts	company	export	cts	dlrs	dollar	bpd	**february**
currency	**corp**	**government**	dollar	dollar	**economic**	company	**increase**
dlrs	group	**import**	**gain**	**first**	**exchange**	**crude**	**january**
dollar	**inc**	japan	**include**	japan	**foreign**	group	japan
loss	market	**japanese**	loss	loss	japan	market	last
mln	**may**	last	market	mark	mark	offer	mark
oper	much	market	net	net	market	oil	offer
p	**new**	much	oper	oil	much	**opec**	pct
pct	offer	**official**	profit	pct	offer	price	profit
profit	pct	pct	qtr	profit	rate	**production**	quarter
qtr	price	say	**sale**	quarter	say	quarter	rate
rate	profit	share	**shr**	rate	share	say	**report**
stg	say	**state**	trade	share	stock	share	**rise**
trade	share	trade	us	stg	us	trade	say
us	stock	us	vs	vs	year	vs	**surplus**
vs	trade	year	year	year	**yen**	year	year

이렇게 선택된 로이터-21578 데이터를 문서-단어행렬(document term matrix; DTM)로 변환하였으며 일반적인 빈도 기준 DTM (TF DTM)과 함께 단어빈도-역문서빈도(term frequency inverse document frequency; TF-IDF) 방식의 DTM (TF-IDF DTM)도 고려하였다. TF-IDF 방식의 경우 모든 문서에서 빈번하게 사용되는 일반적인 의미를 가진 단어의 가중치를 줄이고 특정 문서에서만 사용되는 단어의 가중치를 높임으로써 각 문서의 특징을 더 잘 설명해줄 수 있다는 장점이 있다. TF-IDF 방식의 DTM에서는

$$\text{tf-idf}_{ij} = \text{tf}_{ij} \times \text{idf}_j \tag{3.1}$$

와 같은 방법으로 단어의 출현빈도를 재평가한다. 식 (3.1)에서 tf_{ij}는 i 번째 문서에서 j 번째 단어가 관찰된 빈도 x의 함수로 여러 형태로 정의될 수 있는데 이 논문에서는 $x > 0$일 때 $\text{tf}_{ij} = 1 + \log x$, $x = 0$일 때 $\text{tf}_{ij} = 0$으로 정의하였다. idf_j는 j 번째 단어가 포함된 문서 수의 역수의 함수로 여러 정의 중 $\text{idf}_j = \log(n/\text{df}_j)$를 사용하였다. 여기서 n은 전체 문서의 수, df_j는 전체 문서 중 j 번째 단어를 포함하고 있는 문서의 수에 해당된다. TF-IDF 방식의 DTM과 관련된 더 자세한 내용은 Jeong 등 (2019) 또는 Chang 등 (2020)을 참조할 수 있다.

로이터-21578 데이터의 DTM은 TF-IDF를 적용하여 다음과 같이 변환된다. 로이터-21578 데이터의 각 문서는 기사이므로 기사 내용과 관련된 사람들의 언급을 자주 인용하고 이에 따라 인용문을 나타내는 "say" 등의 단어가 자주 사용된다. 분석에 사용된 로이터-21578 데이터의 기사 3,148건 중 "say"라는 단어가 사용된 기사는 2,493건으로 전체 기사의 79.2%에 해당된다. 이렇게 "say"라는 단어는 여러 문서에 자주 사용되지만 각 기사의 주제를 파악하는 데는 크게 도움이 되지 않으므로 가중치를 작게 평가하는 것이 바람직하다. 반면, "OPEC"이라는 단어는 전체 기사 중 109건의 기사에만 사용된 단어로 빈도는 낮지만 원유와 관련된 기사에만 자주 사용되기 때문에 해당 문서의 주제를 파악하는 데 유용하다. 따라서 "OPEC"이라는 단어는 상대적으로 가중치를 높게 평가하는 것이 바람직하다. 이 연구에서 빈도수 기준 DTM을 TF-IDF 방식의 DTM으로 전환했을 때 "say"의 가중치는 0.233으로, "OPEC"의 가중치는 3.363으로 변환되었다.

Table 4: Selected words for TF-IDF DTM (20 words for each principal component)

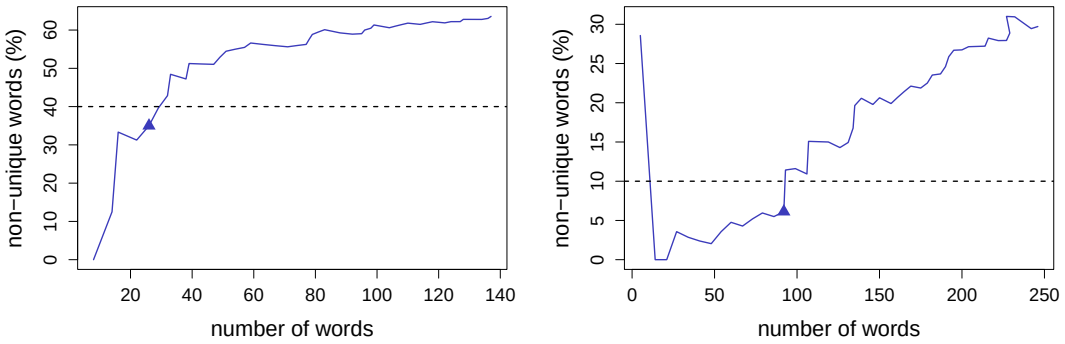
PC1	PC2	PC3	PC4	PC5	PC6	PC7
country	**acquire**	analyst	analyst	**accord**	**account**	**american**
deficit	**acquisition**	**avg**	**barrel**	**baker**	analyst	analyst
ec	analyst	bank	**bpd**	bank	**billion**	can
economic	**bid**	can	can	**central**	**business**	**firm**
export	**board**	**cts**	**crude**	**currency**	**cost**	**go**
foreign	**cash**	foreign	**day**	**dealer**	**earnings**	government
gatt	**common**	**gain**	**demand**	**dollar**	**expect**	**industry**
government	**company**	government	**ecuador**	economic	**fall**	**japan**
import	**file**	**loss**	**export**	**economist**	**first**	**japanese**
minister	**group**	**mths**	**level**	**economy**	**growth**	market
official	**inc**	much	much	**german**	**income**	**may**
policy	**merger**	**net**	**oil**	**germany**	**increase**	much
reagan	**offer**	**nine**	**opec**	**intervention**	**low**	**new**
state	**security**	**oper**	**output**	**mark**	pct	official
surplus	**share**	pct	**pipeline**	market	**profit**	reagan
tariff	**shareholder**	**qtr**	**price**	**monetary**	**quarter**	**sell**
trade	**stock**	**rev**	**production**	**paris**	**report**	**take**
unite	**takeover**	**shr**	rate	rate	**result**	**time**
us	**tender**	us	**saudi**	**west**	**rise**	us
world	**usair**	**vs**	**supply**	**yen**	**year**	**week**

3.2. 희소주성분분석을 이용한 단어 선택

Figure 2의 (a)와 (b)는 단어출현빈도를 기준으로 작성된 DTM에 대해 주성분분석을 적용하였을 때 주성분 수를 구하는 과정을, (c)와 (d)는 TF-IDF를 기준으로 작성된 DTM에 대해 주성분분석을 적용하였을 때 주성분 수를 구하는 과정을 보여준다. Figure 2의 (a)와 (c)는 스크리 그래프를, (b)와 (d)는 인접한 주성분 값의 차이 $d_i - d_{i+1}$ ($i = 1, 2, \ldots, p-1$)를 나타낸다. Figure 2(b)와 (d)에서 i가 커짐에 따라 특잇값의 차이는 0에 가까워짐을 알 수 있다. 특잇값의 차이가 크지 않게 되는 점, 즉 Figure 2(b)에서는 특잇값 차이가 0.2 이하, Figure 2(d)에서는 기울기 차이가 0.1 이하가 되도록 주성분의 수 q를 정하였으며 각각의 경우는 스크리 그래프 (a)와 (c)에서 특잇값 2.3, 2.4에 해당된다. 이렇게 선택된 빈도수 기준 DTM과 TF-IDF 방식 DTM의 주성분 수는 각각 8개와 7개이다.

Tables 3과 4에는 희소주성분분석에서 빈도수 기준 DTM과 TF-IDF 기반의 DTM에서 각각 8개와 7개의 주성분을 선택한 후 각 주성분별로 선형관계식의 계수가 0이 아닌 단어를 20개씩으로 지정하였을 때 선택된 단어들이 기록되어 있다. 표에서 굵은 글자체로 표현되어 있는 단어들은 중복되지 않은 단어이고 굵은 글자로 표현되지 않은 단어들은 두 개 이상의 주성분에서 중복 선택된 단어들에 해당된다. Table 3에서 고유한 단어 수는 71개이고 Table 4에서 고유한 단어 수는 119개이다.

이렇게 TF-IDF 방식의 DTM을 사용하는 경우 상대적으로 많은 단어들이 다른 주성분에서 중복되지 않는 고유한 단어들로 구성되어 있으므로 TF-IDF 방식 DTM의 주성분들이 더 잘 해석되는 반면, 빈도수 기준 DTM의 주성분들은 상대적으로 해석하기 어렵다. 예를 들어 Table 3의 PC1은 은행(bank), 배럴당 유가(bpd), 통화(currency), 이익(profit)과 손해(loss), 무역(trade) 등 서로 다른 주제에 해당되는 다양한 단어들이 섞여 있어 주성분의 의미에 대한 해석이 쉽지 않다. 반면, Table 4에서 PC1의 경우 무역과 관련된 의미를 가지는 주성분, PC2는 기업인수, PC3은 손익, PC4는 원유, PC5는 통화, PC6은 기업수익, PC7은 미일 무역과 관련된 것으로 비교적 명확하게 해석할 수 있다.



(a) TF DTM (b) TF-IDF DTM

Figure 3: *Proportions of non-unique words.*

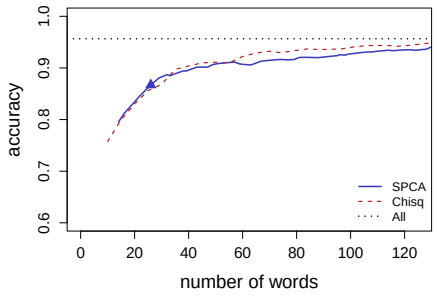
또, 여러 주성분에서 중복되어 선택되는 단어들이 많으면 각 주성분에서 선택되는 단어 수를 늘리더라도 추가로 얻어지는 정보는 많지 않다. 따라서 2절의 단어선택 절차에 제시된 바와 같이 각 주성분에서 중복되는 단어가 많이 나타나지 않도록 각 주성분별 단어 수를 결정하는 방법을 생각해 볼 수 있다. Figure 3은 빈도수 기준 DTM과 TF-IDF 방식의 DTM에 희소주성분분석을 적용하였을 때 선택된 단어들 중 중복 선택된 단어의 비중을 나타낸 그림이다. 빈도수 기준 DTM과 TF-IDF 방식의 DTM 모두 단어 수가 증가함에 따라 중복되는 단어 비중이 점차 증가하는 경향이 있음을 확인할 수 있다. 단어 선택 절차에서 제시된 것처럼 중복된 단어 비중이 큰 폭으로 증가하기 전에 멈추는 경우 빈도수 기준 DTM에서는 26개의 단어, TF-IDF 기준 DTM에서는 92개의 단어가 선택된다. 다만, Figure 3의 (a)와 (b)는 각 주성분에서 선택된 단어 수 m 값이 작을 때는 일부 상이한 형태를 보였다. 즉, TF DTM에서는 단어 수가 증가하면서 중복된 단어의 비중이 높아지는 경향을 보인 반면, TF-IDF DTM에서는 $m = 1$인 경우에 7개의 주성분 중 2개의 주성분에서 동일한 단어가 선택되고 $m = 2$인 경우에는 모든 단어들이 중복되지 않아 중복된 단어의 비중이 28.6%에서 0.0%로 축소되었다.

3.3. 선택된 단어를 이용한 분류분석

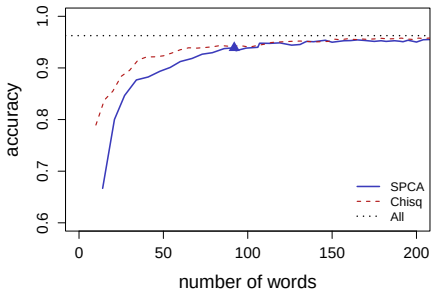
전체 단어 중 일부만 선택하여 분류모형을 적용하였을 때 이들 범주를 정확하게 예측할 수 있다면 선택한 단어들이 텍스트데이터의 범주에 대한 충분한 정보를 보존하고 있다고 간주할 수 있다. 이 절에서는 앞에서 제안한 방식으로 선택된 단어들이 텍스트데이터의 정보를 충분히 포함하고 있는지 확인하기 위해 로이터-21578 데이터에서 선택된 단어들을 이용하여 분류분석을 실시해본다. 이 연구에서 사용한 로이터 기사 데이터는 모두 다섯 개의 범주에 해당되는 기사들로 구성되어 있다.

로이터 데이터의 분류분석을 위해 지지벡터기계(support vector machine; SVM)와 k-최근접이웃 알고리즘을 고려해보았다. SVM의 경우 설명변수가 많은 고차원 데이터에 대해서도 잘 작동하는 반면, KNN 알고리즘의 경우 설명변수가 너무 많은 경우 계산에 많은 시간이 소요될 뿐만 아니라 잡음이 누적되어 분류 정확성이 오히려 떨어질 수 있음이 알려져 있다.

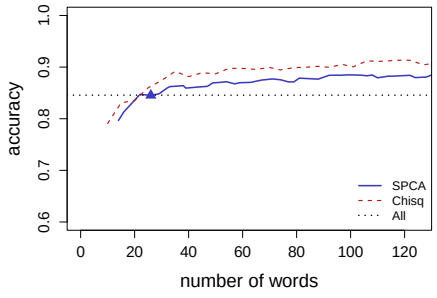
분류분석 결과는 Figure 4에 제시되어 있다. Figure 4에서 점선으로 표현된 수평선은 단어선택 과정 없이 모든 단어를 분류에 사용하였을 때 얻어지는 결과이고 파란색 실선은 희소주성분분석을 이용하여 단어를 선택하였을 때 선택된 단어 수별 분류 정확도를 나타내며 선택된 단어 수가 x축에, 분류분석 결과의 정확도(accuracy)가 y축에 표현되어 있다. 붉은색 파선(dashed line)은 범주에 대한 정보를 활용하여 카이제곱통계량 값을 기준으로 단어를 선택하였을 때의 분류 정확도에 해당된다. 2절에서 제안한 방법으로 단어를 선택하였을 때의 단어 수와 분류 정확도는 파란색 삼각형으로 표현하였다. Figure 4의 왼쪽 그림들은 빈도수 기준의



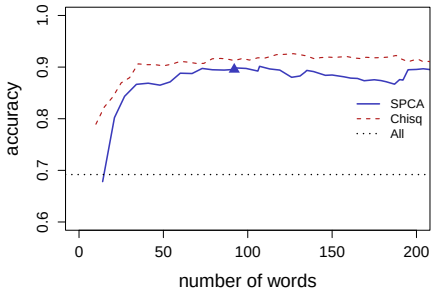
(a) SVM for TF DTM



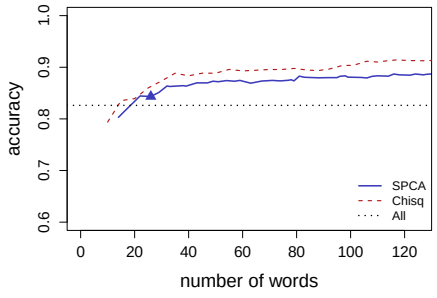
(b) SVM for TF-IDF DTM



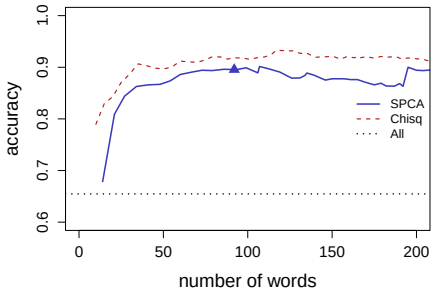
(c) 3-NN for TF DTM



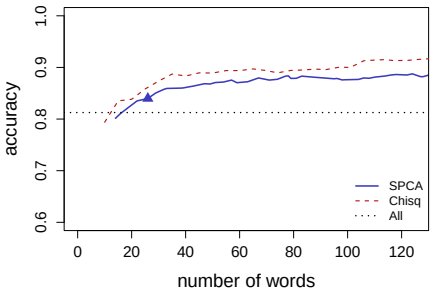
(d) 3-NN for TF-IDF DTM



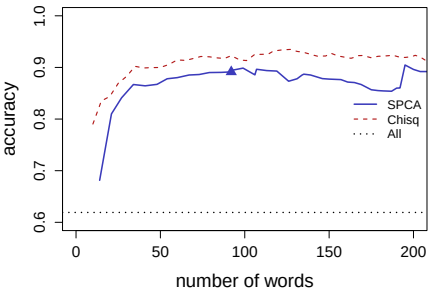
(e) 5-NN for TF DTM



(f) 5-NN for TF-IDF DTM



(g) 7-NN for TF DTM



(h) 7-NN for TF-IDF DTM

Figure 4: *Classification accuracy.*

Table 5: Classification accuracy by the PC scores

	SVM	3-NN	5-NN	7-NN
TF DTM	0.783	0.848	0.830	0.814
TF-IDF DTM	0.834	0.692	0.655	0.619

DTM을 이용하였을 때, 오른쪽 그림들은 TF-IDF 기준의 DTM을 이용하였을 때의 분류 결과를 보여준다.

SVM의 조절모수에 해당되는 비용(cost) C는 교차검증을 통해 0.01로 지정하였다. SVM을 적용하였을 때의 분류 결과는 Figure 4의 (a)와 (b)에서 확인할 수 있다. 이 연구에서는 분류 성능을 측정하기 위해 전체 분류대상 문서 중 정확하게 분류된 문서의 비중을 나타내는 정확도(accuracy)를 이용하였다. Figure 4의 (a)와 (b)에서 SVM을 적용하였을 때는 선택되는 단어 수가 증가할수록 분류 성능이 개선되는 공통점이 있는 것을 알 수 있다. 또, TF-IDF 기준의 DTM에서 소수의 단어만을 선택한 경우를 제외하면 희소주성분분석으로 단어를 선택하였을 때와 카이제곱통계량을 이용하여 단어를 선택하였을 때 분류 성능에 큰 차이가 없었다. 한편, 이 연구에서 제안한 방식으로 TF-IDF 방식 DTM에서 92개의 단어를 선택하고 SVM을 적용하여 얻은 분류 정확도는 0.949로 모든 단어를 사용하여 분류하였을 때의 분류 정확도 0.963과 큰 차이가 없었다. 반면, 일반적인 빈도수 기준 DTM에서 26개의 단어를 선택하였을 때는 SVM을 적용하여 얻은 분류 정확도가 0.867로 단어선택 절차 없이 모든 단어를 분류에 이용하였을 때의 정확도 0.956과 상당한 차이를 보였다. 즉, 2절에서 제안한 절차는 TF-IDF 방식의 DTM에서 더 잘 작동하는 것을 알 수 있다.

다만, 이러한 차이는 제안된 절차를 적용하였을 때 TF-IDF DTM에서 더 많은 단어가 선택되었기 때문으로 볼 수 있다. TF 기반 DTM에서 26개의 단어를 선택한 것과 비슷하게 TF-IDF 기반 DTM에서 27개의 단어를 선택한 경우 정확도는 0.847로, TF-IDF 기반 DTM에서 92개의 단어를 선택한 것과 같이 TF 기반 DTM에서 92개의 단어를 선택한 경우 정확도는 0.922로 나타났다. 즉, 제안된 절차에 따라 단어를 선택하였을 때 확인된 분류 성능의 차이는 선택된 단어 수가 다른 데 따른 것으로 볼 수 있으며 단어 수가 대등할 때는 분류 성능에 큰 차이가 없는 것으로 나타났다.

KNN을 적용하였을 때의 분류 결과는 Figure 4의 (c)–(h)에서 살펴 볼 수 있다. KNN 분류 결과는 근접한 이웃의 수 k 값의 선택에 따라 달라질 수 있으므로 k가 3, 5, 7인 세 가지 경우를 고려하였으며 (c)와 (d)는 k가 3일 때, (e)와 (f)는 k가 5일 때, (g)와 (h)는 k가 7일 때의 분류 결과에 해당된다. 그림에서 확인할 수 있는 바와 같이 SVM을 적용한 경우와 달리 KNN을 적용하였을 때는 모든 단어를 사용하였을 때에 비해 일부 단어만 선택하여 분류기에 입력하였을 때 오히려 분류 정확도가 높게 나타나는 경향을 보였다. 특히 2절에서 제안한 절차를 이용하여 TF-IDF 방식의 DTM에서 단어를 선택할 때 분류 결과가 더 우수한 것을 확인할 수 있다. 즉, TF-IDF 방식의 DTM에서는 전체 단어를 KNN 분류기에 입력하는 경우보다 제안된 절차로 일부 단어만 선택하여 KNN 분류기에 입력할 때 분류 정확도가 크게 향상됨을 알 수 있었다. 다만, 빈도수 기준의 DTM을 이용하였을 때는 전체 단어를 사용한 경우와 제안된 절차로 선택한 단어를 입력한 경우의 분류 정확도 차이가 상대적으로 작았다. 하지만, 이 경우에도 전체 단어를 사용하는 경우에 비해 해석이 쉽고, 계산 시간도 크게 단축되므로 단어선택 절차가 의미를 가진다고 볼 수 있다.

한편, Table 5에는 주성분점수(PC scores)를 이용한 분류분석 결과가 제시되어 있다. 이 표에서 SVM을 주성분점수에 적용하여 분류한 결과보다 제안된 절차를 통해 단어의 원형을 보존한 상태에서 분류한 결과가 더 나은 것을 확인할 수 있다. 각 주성분점수는 모든 단어들의 선형결합으로 생성되었음에도 불구하고 희소주성분분석을 통해 구한 분류 정확도가 더 높은 것은 제안된 절차가 로이터 기사 데이터에 대해 잘 작동함을 보여주는 것으로 해석할 수 있다.

4. 결론 및 토의

희소주성분분석은 벌점이 부여된 최소제곱법 중 하나인 엘라스틱넷을 사용하여 유용하지 않은 주성분을 제거하고 각 주성분에서도 유용한 변수만을 이용하여 주성분을 표현하는 선형식을 만들어낸다. 따라서 희소주성분분석을 이용하여 많은 변수로 이루어진 데이터를 소수의 변수만으로 요약하는 과정을 구현할 수 있다. 이 논문에서는 희소주성분분석의 이러한 특징을 이용하여 텍스트데이터에서 단어를 선택하는 절차를 제안하였다. 이렇게 희소주성분분석을 이용하여 단어를 선택하는 과정을 통해 중요도가 낮은 단어를 제거함으로써 텍스트데이터의 분류 정확성은 유지하면서 데이터의 차원을 축소할 수 있음을 확인하였다. 특히 이렇게 차원을 축소함으로써 KNN 등 고차원 데이터 분석에서 분류 정확도가 저하되는 분류기의 분류 성능을 개선할 수 있음을 알 수 있었다.

또한, 주성분분석의 각 주성분들은 서로 직교하므로 Jang 등 (2022)에서 잠재디리클레할당(LDA)을 통해 구한 토픽들에 비해 보다 간결하게 텍스트 데이터를 요약할 수 있을 것으로 기대할 수 있다. 실제로 Jang 등 (2022)에서는 19개의 토픽으로 로이터 기사 데이터를 요약하는 것이 적절한 것으로 판단된 반면, 이 연구에서는 7~8개의 주성분으로 로이터 데이터를 요약할 수 있는 것으로 확인되었다.

한편, 희소주성분분석을 이용하여 단어를 선택하기 위해서는 적절한 주성분의 수와 주성분별로 중요도가 높은 단어 수를 결정하는 과정이 필요하며 이 연구에서는 스크리 그림과 중복 선택된 단어 비중을 이용하는 방법을 제안하였다. 다만, 이 연구에서 제안된 방식의 경우 단어 수 결정과정에서 주관적인 판단을 완전히 배제하기는 쉽지 않다. 이와 같은 측면에서 교차검증 과정 등을 통해 주성분 수와 계수가 0이 아닌 단어 수를 결정하는 방식도 고려해볼 수 있다. 희소주성분분석을 위해 사용하는 엘라스틱넷의 경우 목적함수가 볼록함수이므로 수치적인 방법으로 정확한 해를 구할 수 있다는 장점이 있다. 하지만 실제 데이터 분석에서 텍스트데이터와 같이 데이터의 크기가 큰 경우에는 수치해를 구하는 과정에서 많은 연산 시간이 소요되기 때문에 교차검증을 적용하여 단어 수를 선택하기 어려웠다. 향후 보다 객관적인 주성분 수 결정 및 선형식에서의 단어 수 선택 과정과 엘라스틱넷의 계산 과정 효율화를 위한 연구가 필요한 것으로 판단된다.

References

Chang Y, Hwang H, and Son W (2020). *Unstructured Data Analysis*, KNOU Press, Seoul.

Chen J, Huang H, Tian S, and Qu Y (2009). Feature selection for text classification with Naive Bayes, *Expert Systems with Applications*, **36**, 5432–5435.

Forman G (2003). An extensive empirical study of feature selection metrics for text classification, *Journal of Machine Learning Research*, **3**, 1289–1305.

Jang W, Kim YE, and Son W (2022). Feature selection for text data via topic modeling, *The Korean Journal of Applied Statistics*, **35**, 739–754.

Jeong HY, Shin SM, and Choi YS (2019). Comparison of term weighting schemes for document classification, *The Korean Journal of Applied Statistics*, **32**, 265–276.

Jolliffe IT (2002). *Principal Component Analysis* (2nd ed), Springer, New York.

Manning C and Schutze H (1999). *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, Massachusetts.

Mladenić D and Grobelnik M (1999). Feature selection for unbalanced class distribution and naïve bayes. In *Proceedings of the 16th International Conference on Machine Learning (ICML)*, Bled, Slovenia, 258–267.

Yang Y and Pedersen JO (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning (ICML)*, Nashville, TN, 412–420.

Zou H, Hastie T, and Tibshirani R (2006). Sparse principal component analysis, *Journal of Computational and*

Graphical Statistics, **15**, 265–286.

Received February 24, 2023; Revised May 2, 2023; Accepted May 27, 2023

희소주성분분석을 이용한 텍스트데이터의 단어선택

손원[1,a]

[a]단국대학교 정보통계학과

요 약

텍스트데이터는 일반적으로 많은 단어로 이루어져 있다. 텍스트데이터와 같이 많은 변수로 구성된 데이터의 경우 과적합 등의 문제로 분석에 있어서의 정확성이 떨어지고, 계산과정에서의 효율성에도 문제가 발생하는 경우를 흔히 볼 수 있다. 이렇게 변수가 많은 데이터를 분석하기 위해 특징선택, 특징추출 등의 차원축소 기법이 자주 사용되고 있다. 희소주성분분석은 벌점이 부여된 최소제곱법 중 하나로 엘라스틱넷 형태의 목적함수를 사용하여 유용하지 않은 주성분을 제거하고 각 주성분에서도 중요도가 큰 변수만 식별해내기 위해 활용되고 있다. 이 연구에서는 희소주성분분석을 이용하여 많은 변수를 가진 텍스트데이터를 소수의 변수만으로 요약하는 절차를 제안한다. 이러한 절차를 실제 데이터에 적용한 결과, 희소주성분분석을 이용하여 단어를 선택하는 과정을 통해 목표변수에 대한 정보를 이용하지 않고도 유용성이 낮은 단어를 제거하여 텍스트데이터의 분류 정확성은 유지하면서 데이터의 차원을 축소할 수 있음을 확인하였다. 특히 차원축소를 통해 고차원 데이터 분석에서 분류 정확도가 저하되는 KNN 분류기 등의 분류 성능을 개선할 수 있음을 알 수 있었다.

주요용어: 변수선택, 엘라스틱넷, 제약이 있는 회귀모형, 희소주성분분석

[1](16890) 경기도 용인시 수지구 죽전로 152, 단국대학교 정보통계학과. E-mail: son.won@dankook.ac.kr