The Korean Journal of Applied Statistics
2022, Vol. 35, No. 6, 739–754

DOI: https://doi.org/10.5351/KJAS.2022.35.6.739
Print ISSN1225-066X / Online ISSN 2383-5818

Feature selection for text data via topic modeling

Woosol Jang[a], Ye Eun Kim[a], Won Son[1,a]

[a]Department of Applied Statistics, Dankook University

Abstract

Usually, text data consists of many variables, and some of them are closely correlated. Such multi-collinearity often results in inefficient or inaccurate statistical analysis. For supervised learning, one can select features by examining the relationship between target variables and explanatory variables. On the other hand, for unsupervised learning, since target variables are absent, one cannot use such a feature selection procedure as in supervised learning. In this study, we propose a word selection procedure that employs topic models to find latent topics. We substitute topics for the target variables and select terms which show high relevance for each topic. Applying the procedure to real data, we found that the proposed word selection procedure can give clear topic interpretation by removing high-frequency words prevalent in various topics. In addition, we observed that, by applying the selected variables to the classifiers such as naïve Bayes classifiers and support vector machines, the proposed feature selection procedure gives results comparable to those obtained by using class label information.

Keywords: feature selection, latent Dirichlet allocation, text data, topic model, unsupervised learning

1. 서론

텍스트 데이터는 일반적으로 많은 다양한 단어들로 이루어져 있다. 텍스트 데이터에는 수만 개에서 수십만 개의 고유한 단어들이 등장하기도 한다. 따라서 하나의 고유한 단어를 하나의 변수로 볼 때 텍스트 데이터는 변수의 개수가 많은 데이터에 해당된다. 많은 단어들 중 일부는 전체 맥락을 파악하는 데 중요한 의미를 내포하고 있지만 상당수 단어들은 문법적 역할을 하거나 문서의 전반적인 의미를 파악하는 데 크게 도움이 되지 않는다. 이렇게 문서의 전반적 의미와 관련성이 낮은 잡음(noise)에 해당되는 변수들이 많이 포함되어 있는 경우 통계 분석의 정확성, 효율성 등이 떨어질 수 있으므로 정확하고 효율적인 분석을 위해 변수 선택(feature selection) 또는 변수 추출(feature extraction) 등의 방법이 활용되기도 한다. 변수 선택에서는 변수의 원형을 보존하면서 분석에 포함시킬 변수와 제외할 변수를 선택하는 반면, 변수 추출에서는 주성분 분석에서처럼 변수의 선형 조합 등을 통해 새로운 변수를 생성하여 분석에 사용할 수 있다.

목표 변수(target variable)가 있는 지도 학습(supervised learning)에서는 단어 선택을 위해 목표 변수와 단어의 출현 빈도 사이의 연관성을 이용할 수 있다. 예를 들어 교육을 주제로 하는 문서를 식별하기 위해서는 교육과 관련된 문서에 자주 등장하는 학교, 학생, 학습 등의 단어를, 예술과 관련된 문서를 식별하기 위해서는 미술관, 작품, 전시회, 연주 등의 단어를 이용하여 각 주제에 해당되는 문서들을 분류해낼 수 있다. 하지만, 비지도 학습에서는 지도 학습에서와 달리 단어 선택의 기준으로 삼을 수 있는 목표 변수에 대한 정보가 없기 때문에 지도 학습에서와 같은 단어 선택 방법을 적용하기 위해서는 추가적인 절차가 필요하다.

[1] Corresponding author: 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 2022 / journal homepage: http://kjas.or.kr

이 연구에서는 목표 변수에 대한 정보가 없는 비지도 학습에서 단어를 선택하기 위해 토픽 모형을 이용하는 절차를 제안한다. 토픽 모형에서 토픽이란 함께 모여서 특정한 의미를 나타내는 단어들의 모임을 말한다 (Boyd-Graber 등, 2017). 즉, 토픽 모형은 텍스트 데이터의 주제를 파악하기 위해 사용되는 통계적 분석 방법 (Blei, 2012)으로 볼 수 있다. 문서의 주제를 파악하기 위해 다양한 토픽 모형이 제안되어 왔는데 대표적인 것으로는 LSA (latent semantic analysis) 또는 LSI (latent semantic indexing)로 지칭되는 잠재 의미 분석 (Deerwester 등, 1990), pLSA (probabilistic latent semantic analysis)로 지칭되는 확률적 잠재 의미 분석 (Hofmann, 1999), LDA (latent Dirichlet allocation)라 불리는 잠재 디리클레 할당 (Blei 등, 2003) 등이 있다. LSA의 경우 주성분 분석에서와 같이 문서-단어 행렬(document-term matrix)을 분해하였을 때 하나의 고윳값과 고유 벡터에 해당되는 성분을 하나의 토픽으로 간주한다. 반면, pLSA와 LDA는 텍스트 데이터에 대한 확률 모형을 가정하고 이 확률 모형을 기반으로 추정된 모수들을 이용하여 토픽을 찾는 방식이라는 점에서 LSA와 차이가 있다.

대표적인 확률적 토픽 모형인 LDA를 이용하면 각 토픽별로 단어들의 평균적인 출현 빈도를 확인할 수 있다. LDA에서는 각 토픽별로 단어의 출현 빈도가 다항 분포에 의해 결정되는 것으로 가정하고 있으므로 다항 분포의 모수들을 이용하여 중요한 단어와 그렇지 않은 단어를 선별할 수 있다. 즉, z라는 토픽에서 다항 분포의 모수 추정값이 큰 단어 w는 토픽 z에서 출현 빈도가 높은 단어이다. 하지만 출현 빈도가 높은 단어들 중에는 여러 토픽에 공통적으로 자주 사용되어 토픽들을 구분하는 데 큰 도움이 되지 않는 단어들도 포함되어 있을 수 있다. 따라서 이 연구에서는 먼저 토픽 모형을 통해 각 토픽에서 자주 사용되는 주요 단어들을 선택하고 이 단어들 중에서 토픽을 식별하는 데 유용한 단어들만 다시 선별하는 방식을 제안한다.

이 논문은 다음과 같이 구성된다. 먼저 2절에서는 대표적인 토픽 모형인 LDA를 중심으로 확률적 토픽 모형에 대한 선행 연구 결과를 정리해본다. 3절에서는 카이제곱 통계량을 이용한 단어 선택 방법에 대해 간략히 살펴본 후 LDA를 이용하여 토픽별로 자주 사용되는 단어들을 식별하고 카이제곱 통계량을 이용하여 각 토픽을 구분하는 데 유용한 단어들을 재선별하는 방법을 제안한다. 4절에서는 3절에서 제시된 단어 선택 방법을 실제 데이터에 적용하여 유용성을 확인해본다. 마지막으로 5절에서는 결과를 정리한다.

2. 확률적 토픽 모형

토픽 모형은 텍스트 데이터에 잠재되어 있는 주제를 확인하기 위한 통계적 분석 방법으로 널리 활용되고 있다. 토픽 모형에는 LSA와 같은 비확률적 토픽 모형과 pLSA, LDA 등의 확률적 토픽 모형이 있는데 이 절에서는 가장 많이 사용되는 토픽 모형 중 하나인 LDA를 중심으로 확률적 토픽 모형에 대해 살펴본다.

이하에서는 관찰된 순서를 가리키는 인덱스는 아래 첨자로, 서로 다른 고유한 값을 가리키는 인덱스는 윗첨자로 표현하기로 한다. 예를 들어 단어 수가 N인 문서에서 n 번째로 관찰된 단어는 w_n으로 표현하고 서로 다른 V개의 고유한 단어들 중 v 번째 단어는 w^v로 표현한다. 따라서 문서 d의 n 번째 관측값이 단어 w^v인 경우 $w_n = w^v$로 표현하고 간략하게 $w_n^v = 1$로 표현하기로 한다.

2.1. Probabilistic latent semantic analysis (pLSA)

확률적 토픽 모형에서는 각 단어들이 다항 분포 등 특정 확률 분포로부터 생성되었다고 가정한다. 예를 들어 가장 단순한 확률적 토픽 모형인 유니그램(unigram) 모형에서는 각 문서가 하나의 다항 분포에서 추출된 단어들로부터 생성된다고 가정한다. 유니그램 모형에서는 나이브 베이즈(naïve Bayes) 분류 모형에서와 같이 각 단어들은 서로 독립이라고 가정하므로 단어의 선후 관계, 즉 맥락과 관련된 정보는 고려하지 않는다. N개의 단어로 이루어진 문서 d에서 n 번째로 관찰된 단어를 w_n이라 할 때 문서 벡터 $d = (w_1, w_2, \ldots, w_N)$은

$$p(d) = \prod_{n=1}^{N} p(w_n)$$

과 같은 확률 모형으로 표현할 수 있다. 하지만, 실제 텍스트 데이터에서는 문서의 주제에 따라 단어의 경험 분포에 큰 차이가 있으므로 유니그램 모형에서와 같이 모든 문서가 하나의 다항 분포로부터 생성될 수 있다고 보는 것은 지나치게 단순한 가정이다.

pLSA (Hoffman, 1999)에서는 단어들의 출현 확률이 토픽에 따라 달라질 수 있다고 가정한다. 즉, pLSA는 하나의 문서에는 다양한 토픽들이 포함될 수 있으며 각 토픽마다 단어들의 분포가 서로 다르다고 가정하는 혼합 모형(mixture model)이다. 유니그램 모형에서는 토픽에 관계없이 $n(n = 1, \ldots, N)$ 번째에 단어 w^v가 관찰될 확률 $p(w_n = w^v)$가 모두 동일한 다항 분포를 따른다고 가정하는 반면, pLSA에서는 단어 w^v가 관찰될 확률이 토픽 z에 따라 달라진다고 가정한다. 즉, 단어 $w_n = w^v$일 확률은 w_n을 생성한 토픽 z_n에 의해 결정되며 조건부 분포 $p(w_n = w^v|z_n)$으로 표현될 수 있다. 한편, pLSA에서는 문서 d에서 각각의 토픽이 사용될 확률도 모형화하여 조건부 분포 $p(z|d)$로 나타내는 계층적 구조를 가정한다. 따라서 문서 d가 주어졌을 때 $w_n = w^v$일 확률을 조건부 분포

$$p\left(w_n = w^v|d\right) = \sum_{k=1}^{K} p\left(z_n^k|d\right) p\left(w_n = w^v|z_n^k\right)$$

로 나타낼 수 있다. 즉, n 번째 단어 $w_n = w_v$가 관찰될 확률은 토픽 z_n^k가 선택될 확률과 토픽 z_n^k가 선택되었을 때 단어 w_v가 관찰될 확률의 곱을 $k = 1, 2, \ldots, K$에 대해 더한 값으로 표현할 수 있다. 이 식에서 n 번째에 관찰된 단어를 생성하는 토픽 $z_n^1, \ldots, z_n^K$는 실제로는 관찰되지 않는 잠재 변수에 해당된다.

2.2. Latent Dirichlet allocation (LDA)

pLSA와 마찬가지로 잠재 디리클레 할당(LDA)도 계층적 구조를 가지는 혼합 모형에 해당된다. LDA에서도 각 문서에 여러 개의 토픽이 포함될 수 있다고 가정한다. 즉, 하나의 문서에서 n 번째로 관찰된 단어 $w_n(n = 1, \ldots, N)$은 각각 K개의 토픽 $z^1, \ldots, z^K$ 중 하나로부터 생성되는 것으로 가정한다는 점에서는 pLSA와 공통점이 있다.

한편, LDA에서는 각 단어를 생성하는 토픽이 다항 분포로부터 결정되며 이 다항 분포의 모수가 디리클레 분포(Dirichlet distribution)에 의해 정해지는 것으로 가정한다는 점에서 pLSA와 차이가 있다. 즉, 토픽을 생성하는 다항 분포의 모수 $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$는 모수가 $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$인 디리클레 분포

$$p\left(\boldsymbol{\theta}|\boldsymbol{\alpha}\right) = \mathrm{Dir}\left(\boldsymbol{\alpha}\right) = \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma\left(\alpha_i\right)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_K^{\alpha_K - 1}$$

로부터 주어지고, $\boldsymbol{\theta}$가 주어졌을 때 토픽 z_n의 확률

$$p\left(z_n|\boldsymbol{\theta}\right)$$

은 다항 분포에 해당된다.

토픽이 K개이고 텍스트 데이터에 포함된 고유한 단어가 V개일 때 각 토픽에 해당되는 다항 분포의 모수들은 $K \times V$ 행렬 $\boldsymbol{\beta}$로 표현할 수 있다. 즉, 행렬 $\boldsymbol{\beta}$의 (k, v) 원소는 k 번째 토픽 z^k에서 v 번째 단어 w^v가 관찰될 확률 $\beta_{kv} = p(w^v = 1|z^k = 1)$로 정의된다. 토픽의 분포를 결정하는 디리클레 분포의 모수 $\boldsymbol{\alpha}$와 각 토픽별 단어의 출현 빈도를 결정하는 다항 분포의 모수 행렬 $\boldsymbol{\beta}$는 전체 텍스트 데이터의 모든 문서들에 공통적으로 적용되는 모수이다. 반면, 토픽의 분포를 결정하는 다항 분포의 모수 $\boldsymbol{\theta}$는 하나의 문서 내에서는 동일한 값을 가지지만 각 문서별로는 서로 다른 값을 가지고, N개의 단어를 포함하고 있는 문서에서 $n(n = 1, \ldots, N)$ 번째에 관찰되는 단어 w_n과 해당 단어를 생성하는 토픽 z_n은 매번 랜덤으로 결정되는 확률 변수이다.

토픽 z_n과 행렬 $\boldsymbol{\beta}$가 주어져 있을 때 단어 w_n의 분포는

$$p\left(w_n|z_n, \boldsymbol{\beta}\right)$$

로 표현할 수 있다. 지금까지의 결과를 정리하면 전체 텍스트 데이터에서 문서 d가 관찰될 확률은

$$\begin{aligned} p(d|\boldsymbol{\alpha},\boldsymbol{\beta}) &= \int p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{n=1}^{N} \sum_{z_n} p(z_n|\boldsymbol{\theta})\, p(w_n|z_n,\boldsymbol{\beta})\; d\boldsymbol{\theta}, \\ &= \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \int \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \prod_{n=1}^{N} \sum_{i=1}^{K} \prod_{v=1}^{V} \left(\theta_i \beta_{ij}\right)^{w_n^v} d\boldsymbol{\theta} \end{aligned}$$

로 주어진다.

3. 비지도 학습에서의 단어 선택

3.1. 카이제곱 통계량을 이용한 단어 선택

변수가 많은 데이터에 선형 모형을 적용할 때 변수 선택 과정을 거치는 것과 같이 많은 단어로 이루어진 텍스트 데이터에서도 분석 목적에 적합한 단어를 찾는 단어 선택 과정이 사용되기도 한다. 특히 텍스트 데이터에는 대다수 문서에 공통적으로 사용되는 단어나 극히 일부 문서에만 등장하는 단어들이 많이 포함되어 있는데 이런 단어들의 경우 문서의 의미를 통계적으로 분석하는 데 큰 도움이 되지 않는 경우가 많다. 전자의 예로는 영어의 관사나 우리말의 조사 등을 들 수 있고 후자의 예로는 일부 문서에만 사용되는 전문 용어나 고유 명사들이 있다. 텍스트 데이터의 의미를 분석하는 데 있어서 중요한 단어는 특정 범주와 연관성이 높은 단어라고 볼 수 있으며 이런 연관성을 파악하기 위해 다양한 방법이 제안되었다. 단어 선택을 위한 방법들에 대한 세부적인 내용은 Forman (2003), Son (2020), Mun과 Son (2022) 등을 참조할 수 있다.

대표적인 단어 선택 방법 중 하나로는 카이제곱 통계량을 이용하는 방법을 들 수 있다. 카이제곱 통계량은 문서의 범주와 단어의 출현 여부를 기준으로 작성한 이차원 분할표를 이용하여 문서를 분류하는 데 도움이 되는 단어를 식별하는 지표로 카이제곱 통계량이 높은 단어를 해당 범주와 관련성이 높은 단어로 선택한다. Table 1에서 p_{11}은 범주 C에 해당하는 문서 중 단어 W를 포함하고 있는 문서의 비중을, p_{10}은 범주 C에 해당되는 문서 중 단어 W를 포함하지 않는 문서의 비중을 나타낸다. 마찬가지로 p_{01}은 범주 C에 해당되지 않는 문서 중 단어 W를 포함하고 있는 문서의 비중, p_{00}은 범주 C에 해당하지 않는 문서 중 단어 W를 포함하지 않는 문서의 비중을 나타낸다. 따라서 단어 W가 문서의 범주를 판단하는 데 도움이 된다면 단어 W가 포함된 문서들 중 많은 문서들이 범주 C에 해당되고 단어 W가 포함되지 않은 문서들 중 많은 문서들이 범주 C에 해당되지 않거나, 단어 W가 포함된 문서들 중 많은 문서들이 범주 C에 해당되지 않고 단어 W가 포함되지 않은 문서들 중 많은 문서들이 범주 C에 해당되어야 한다. 즉, $p_{11}p_{00}$과 $p_{10}p_{01}$의 차이가 클 때 단어 W가 문서의 범주를 식별하는 데 도움이 된다. 이렇게 이차원 분할표의 정보를 요약하는 지표로 카이제곱 통계량 이외에도 오즈비(odds ratio), 정보이득(information gain) 등이 있는데 카이제곱 통계량은

$$\chi^2 = \sum_{i,j} \frac{N\left(p_{ij} - \mathrm{E}\left[p_{ij}\right]\right)^2}{\mathrm{E}\left[p_{ij}\right]} \tag{3.1}$$

과 같이 표현된다. 이 식에서 기댓값 $\mathrm{E}[p_{ij}] = p_{i\cdot}p_{\cdot j}$로 정의된다. 식 (3.1)의 카이제곱 통계량은 간단한 계산 과정을 통해

$$\chi^2 = \frac{N(p_{11}p_{00} - p_{10}p_{01})^2}{p_{1\cdot}p_{0\cdot}p_{\cdot 1}p_{\cdot 0}} \tag{3.2}$$

과 같은 형태로 표현할 수도 있다.

이렇게 각 단어 W_i와 범주 C_j에 대한 이차원 분할표와 카이제곱 통계량을 작성하고 카이제곱 통계량 값이 큰 단어를 선택하는 과정을 범주별로 반복하여 유용한 단어들을 선택한다. 한편, 실제 텍스트 데이터 분석에서는 카이제곱 통계량을 적용하여 가설 검정을 하였을 때 많은 단어들에서 귀무 가설이 기각되어 지나치게 많은 단어가 선택되기도 하므로 통계적 가설 검정 절차를 적용하지 않고 주관적으로 설정된 상위 m개의 단어를 선택하는 경우도 있다 (Forman, 2003). 다만 이렇게 주관적으로 단어의 개수를 선택하면 좋은 분석 결과를 얻기 어려울 수 있으므로 다중검정 등 통계적 타당성을 갖춘 단어 개수 선택 방법을 적용하는 것이 바람직한 것으로 보인다.

3.2. LDA를 이용한 단어 선택

위에서 살펴본 바와 같이 지도 학습에서는 목표 변수에 해당되는 각 문서별 범주와 설명 변수에 해당되는 단어 출현 여부에 대한 정보를 이용하여 해당 범주를 잘 대표하는 단어를 식별할 수 있다. 한편, 목표 변수가 주어지지 않는 비지도 학습에서는 문서의 범주에 대한 정보를 사용할 수 없으므로 범주 대신 토픽 모형에서 식별된 각 토픽에 대한 정보를 이용하는 방법을 고려해볼 수 있다. 즉, 각 토픽의 비중이 높은 문서에서의 단어 출현 빈도를 이용하여 단어를 선택하는 방법을 생각해볼 수 있다. 일반적으로 토픽 모형에서는 토픽-단어 행렬 $\boldsymbol{\beta}$를 이용하여 각 토픽별로 출현 빈도가 높은 단어들을 선택하는 경우가 많다. 하지만, 단순한 출현 빈도만 이용하여 단어들을 선택하는 경우 여러 토픽에서 공통적으로 자주 등장하는 단어들로 인해 토픽 사이의 차이를 확인하기 어렵고 통계 분석의 정확성이 떨어질 수 있기 때문에 토픽별로 특징적인 단어들을 다시 선택하는 과정을 추가하는 것이 바람직한 것으로 판단된다. 토픽 모형을 이용한 단어 선택은 다음과 같은 절차로 정리할 수 있다. 먼저 LDA 등 토픽 모형을 이용하여 토픽을 식별한다. 다음으로 텍스트 데이터의 문서들 중 추정된 토픽 z_k의 비중 $\hat{\theta}_k$값이 q 이상인 문서를 토픽 z_k와 관련된 문서로, 비중이 q 미만인 문서를 토픽 z_k와 관련성이 낮은 문서로 간주한다. 여기서 $\hat{\theta}_k$는 LDA 모형에서 추정한 토픽 비중으로 다항 분포의 모수 추정값에 해당된다. 다음으로 각 토픽별로 단어 w의 출현 여부를 확인하여 이차원 분할표를 작성하고 카이제곱 통계량을 계산한다. 마지막으로 각 토픽별로 카이제곱 통계량 값이 큰 m개 단어와 토픽별 단어 출현 빈도를 나타내는 토픽-단어 행렬 $\boldsymbol{\beta}$의 추정값이 큰 m개 단어의 교집합에 포함되는 단어들을 선택한다.

Procedure 1 Feature selection via topic model

1. Find topics for the given text data.
2. For a topic z_k and a document d, if $\hat{\theta}_k \geq q$ then classify d as $d \in z_k$, if not, classify d as $d \notin z_k$.
3. For each topic and word, make a contigency table and evaluate a chi-square statistic.
4. For each topic, select words as the intersection of the top-m features based on chi-square statistics and the top-m features in the topic-word matrix $\boldsymbol{\beta}$.

토픽 모형에서는 적절한 토픽의 개수 K가 얼마인지 결정할 필요가 있다. 토픽 수를 결정하기 위한 다양한 방법이 제안되었는데 대표적인 방법으로는 Blei 등 (2003), Griffiths와 Steyvers (2004), Cao 등 (2009), Arun 등 (2010), Deveaud 등 (2014)을 들 수 있다. Blei 등 (2003)의 퍼플렉시티(perplexisity)와 Griffiths와 Steyvers (2004)가 제안한 방법은 가능도 함수를 기준으로 적절한 토픽의 개수를 결정한다. Cao 등 (2009)은 각 토픽 사이의 코사인 유사도(cosine similarity)를 기준으로, Arun 등 (2010)과 Deveaud 등 (2014)은 토픽-단어 행렬, 문서-토픽 행렬 등을 이용하여 구한 대칭형 쿨백-라이블러 발산(symmetric Kullback-Leibler divergence)을 사용하여 적절한 토픽 수를 선택하는 방법을 제안한 바 있다. 토픽 수 선택과 관련된 보다 자세한 내용은 Lee 등 (2021)을 참조할 수 있다.

한편, 제안된 절차의 과정 2에서는 각 문서가 특정 주제에 해당되는지에 대해 판정하기 위해 $\hat{\theta}_k$값과 q를

Table 1: Proportions of documents classified by documents' category and existence of word W

		Existence of word W		total
		true	false	
Membership of category C	true	p_{11}	p_{10}	$p_{1\cdot}$
	false	p_{01}	p_{00}	$p_{0\cdot}$
total		$p_{\cdot 1}$	$p_{\cdot 0}$	$p_{\cdot\cdot} = 1$

비교하였다. 임계값 q는 0에서 1 사이의 값으로 지정할 수 있는데 1에 가까운 큰 값으로 선택하면 $\hat{\theta}_k$가 큰 경우, 즉 토픽 z_k의 비중이 높은 것으로 추정된 경우에만 해당 문서를 Table 1의 범주 $C = z_k$에 해당되는 것으로 보고, 나머지 문서들은 범주 z_k에 해당되지 않는 것으로 간주하여 이차원 분할표를 작성한다. 이와 반대로 임계값 q가 작은 값으로 선택되는 경우에는 z_k를 포함하여 여러 다양한 토픽들이 혼재되어 있는 문서도 범주 z_k에 해당되는 것으로 간주될 수 있다.

4. 실제 데이터에의 적용

앞에서 제안한 토픽 모형을 이용한 단어 선택 절차를 실제 데이터에 적용하여 단어를 선택하고 단어 선택 절차의 유용성을 확인해본다. 분석 대상 데이터로는 로이터-21578 데이터와 20-뉴스그룹(news groups) 데이터를 사용한다. 로이터-21578 데이터는 로이터(Reuters) 통신사의 기사 약 1만 건을 모아둔 데이터로 약 90개의 많은 범주로 이루어져 있는데 상당수 범주는 적은 수의 기사들로만 이루어져 있으므로 많은 기사를 포함하고 있는 기업 인수(acq), 원유(crude), 수익성(earn), 외환(money-fx), 무역(trade) 등의 다섯 개 범주에 해당되는 기사 3,148건만 선택하였다. 20-뉴스그룹 데이터는 유즈넷(usenet) 게시판의 게시물들을 모아둔 데이터로 20개 범주로 구성되어 있다. 로이터 기사 데이터에서와 마찬가지로 20개 범주 중에서 판매(misc.forsale), 자동차(rec.autos), 야구(rec.sport.baseball), 전자 제품(sci.electronics), 의학(sci.med) 등 다섯 개의 범주에 해당되는 2,198건의 게시물만 분석 대상에 포함한다. 대소문자 전환, 불용어(stopwords) 제거 등의 기본적인 전처리 과정을 거친 후 로이터 기사 데이터는 1,107개, 뉴스그룹 게시물 데이터는 1,331개의 단어를 포함하고 있다.

4.1. 토픽 모형을 이용한 단어 선택

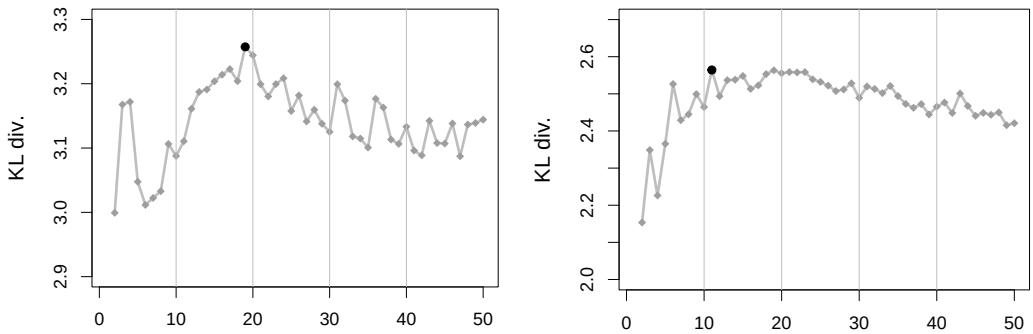
토픽 모형을 이용하여 단어를 선택하기 위해서 먼저 각 데이터별로 적합한 토픽의 수를 결정할 필요가 있다. 이 연구에서는 토픽 수를 선택하기 위해 Deveaud 등 (2014)이 제안한 방식을 이용한다. Deveaud 등 (2014)은 K개의 토픽으로 이루어진 토픽 집합 T_K에서 두 개 토픽의 순서쌍 (i, j)를 선택하고 이 두 토픽들의 쿨백-라이블러 발산의 평균값을 최대화하는 토픽의 개수를 선택할 것을 제안하였다. 즉, Deveaud 등 (2014)이 제안한 절차에서는 쿨백-라이블러 발산 $D(\cdot\|\cdot)$에 대해

$$\hat{K} = \arg\max_{K} \frac{1}{K(K-1)} \sum_{\substack{i,j=1,\ldots,K \\ i\neq j}} D\left(z_i \| z_j\right) \tag{4.1}$$

와 같이 토픽들 사이의 쿨백-라이블러 발산의 평균값을 최대화하는 토픽 수를 선택할 것을 제안한 바 있다. 두 토픽 z_i와 z_j 사이의 쿨백-라이블러 발산 $D(z_i\|z_j)$ 값은 각 토픽의 단어 출현 확률 $P(w|z_i)$와 $P(w|z_j)$를 이용하여

$$D\left(z_i \| z_j\right) = \sum_{w} P(w|z_i) \log \frac{P(w|z_i)}{P(w|z_j)}$$

와 같이 정의할 수 있다. Deveaud 등 (2014)은 각 토픽별로 출현 빈도가 높은 m개의 단어에 한정하여 $D(z_i\|z_j)$ 값을 계산하는 방식을 제안하였다. 식 (4.1)의 쿨백-라이블러 발산값은 R 패키지 `ldatuning`을 이용하여 구할 수 있다.



(a) Reuters (b) 20 News Groups

Figure 1: *Choice of the number of topics.*

Figure 1에는 Deveaud 등 (2014)에서 제안된 방식으로 계산한 쿨백-라이블러 발산값이 표현되어 있다. 로이터 기사 데이터의 경우 토픽 수가 19개일 때 가장 큰 값을 나타내고, 뉴스그룹 게시물 데이터의 경우 토픽 수가 11개일 때 가장 큰 값이 관찰되었다. 따라서, 로이터 기사 데이터의 토픽 수는 19개로, 뉴스그룹 게시물 데이터의 토픽 수는 11개로 지정한다.

이렇게 결정된 토픽 수를 이용하여 LDA 모형을 추정한다. LDA 모형은 R 패키지 `topicmodels`를 이용하여 추정할 수 있다. LDA 모형 추정 결과로부터 각 토픽별 단어 출현 빈도를 결정하는 다항 분포의 모수 행렬 $\boldsymbol{\beta}$를 확인할 수 있다. Table 2에는 각 토픽별 다항 분포의 모수 추정값 중에서 상위 10개씩에 해당되는 단어들이 기록되어 있다. 표에서 로이터 기사 데이터의 경우 “say”, “mln”, “company”, “dlrs” 등의 단어는 여러 토픽에서 등장 확률이 높은 것을 확인할 수 있다. 예를 들어 “say”는 19개의 토픽 중 16개의 토픽에서, “mln”은 7개의 토픽에서, “company”와 “dlrs”는 6개의 토픽에서 출현 빈도 상위 10개의 단어에 포함되었다. 마찬가지로 뉴스그룹 게시물 데이터의 경우에도 11개의 토픽 중 “good”과 “use”가 각각 6개 토픽에서, “can”과 “get”이 각각 5개 토픽에서 출현 확률이 높은 상위 10개의 단어 목록에 포함된 것을 볼 수 있다.

위에서 살펴본 바와 같이 LDA 모형에서 구한 토픽-단어 행렬 $\boldsymbol{\beta}$는 토픽별 단어 출현 확률에 해당되므로 하나의 단어가 여러 개의 토픽에서 출현 확률이 높게 나타날 수 있다. 이렇게 많은 토픽에서 출현 확률이 높은 단어들은 텍스트 데이터의 분석에서 유용하지 않을 수 있으므로 3절에서 제안한 절차를 통해 각 토픽별로 유용한 단어를 선택한다. LDA 모형에서 구한 문서-토픽 행렬로부터 각 문서들이 어떤 토픽 성분을 가지고 있는지 확인할 수 있으며 3.2절에서 제안된 절차의 2를 진행할 수 있다. i 번째 문서에서 j 번째 토픽 성분이 정해진 임계값 q 이상일 때 i 번째 문서가 j 번째 토픽에 해당되는 것으로 판단한다.

임계값 q에 따라 각 토픽별로 선택되는 단어들이 달라질 수 있으므로 q의 선택이 전체 분석 절차에 영향을 미칠 수 있다. q값이 작을 때는 하나의 문서가 여러 토픽으로 할당되고, 반대로 q값이 클 때는 토픽이 할당되지 않는 문서가 다수 관찰될 수 있으므로 적절한 균형점을 찾는 것이 바람직한 것으로 판단하여 이 연구에서는 아래와 같은 방식으로 q값을 설정하였다. Table 3에는 로이터 기사 데이터와 뉴스그룹 데이터에서 임계값 q의 변화에 따라 각 문서에 할당된 토픽의 수가 기록되어 있다. 예를 들어 로이터 데이터에서 $q = 0.4$인 경우 433건의 문서는 모든 토픽의 $\hat{\theta}$값이 0.4 미만이므로 토픽이 할당되지 않았고 2,603건의 문서는 $\hat{\theta}$값이 0.4 이상인 토픽이 하나만 존재하여 하나의 토픽만 할당되었다. 또 $\hat{\theta}$값이 0.4 이상인 토픽이 두 개인 경우도 112건으로 관찰되었다. 이 연구에서는 토픽이 할당되지 않거나 둘 이상의 토픽이 할당된 문서의 수가 상대적으로 적은 값인 $q = 0.4$를 적절한 임계값으로 판단하여 $q = 0.4$를 기준으로 제안된 절차의 과정 2를 진행하였다.

다음으로 이차원 분할표를 작성하고 각 토픽별로 카이제곱 통계량 값이 가장 큰 m개의 단어를 선택하고 토픽-단어 행렬 $\boldsymbol{\beta}$에서 각 토픽별로 등장 확률이 높은 m개 단어와의 교집합을 구하여 단어를 선택한다.

Table 2: Top-10 words of the LDA topics

	Top-10 words									
	1	2	3	4	5	6	7	8	9	10
	<Reuters-21578 Data>									
1	say	company	american	**merger**	can	much	**corp**	analyst	service	new
2	say	**foreign**	**government**	**country**	debt	**export**	**brazil**	year	world	budget
3	**oil**	say	price	**barrel**	**crude**	**opec**	mln	**bpd**	gas	dlrs
4	billion	dlrs	**trade**	mln	february	**surplus**	**january**	**deficit**	year	**export**
5	billion	**profit**	pct	**mark**	say	year	mln	**franc**	rise	group
6	**vs**	mln	dlrs	loss	**cts**	**net**	**shr**	profit	**qtr**	oper
7	dlrs	mln	say	**quarter**	**company**	year	**earnings**	share	first	loss
8	say	group	pct	share	**stake**	**usair**	**security**	stock	buy	company
9	mln	**vs**	**p**	**stg**	profit	billion	**plc**	**pretax**	interest	net
10	say	**ltd**	pct	canadian	company	canada	new	sale	**stake**	**unit**
11	**japan**	say	**japanese**	us	trade	official	market	**semiconductor**	unite	state
12	say	**company**	**inc**	agreement	**board**	dlrs	**gencorp**	also	general	co
13	**offer**	say	**share**	dlrs	**bid**	**tender**	analyst	stock	company	per
14	say	market	**rate**	exchange	**currency**	much	dollar	policy	trader	future
15	bank	mln	loan	**stg**	market	say	**money**	saving	billion	**bill**
16	say	**dollar**	**yen**	**currency**	bank	us	japan	**dealer**	rate	exchange
17	**trade**	say	us	**ec**	**import**	**gatt**	year	country	bill	state
18	pct	say	year	us	**rise**	**growth**	analyst	much	increase	last
19	**share**	say	**stock**	dlrs	company	**common**	pct	**shareholder**	mln	**dividend**
	<20 Newsgroups Data>									
1	one	good	use	can	think	make	**science**	many	result	work
2	**disease**	**medical**	**health**	use	university	report	center	**cancer**	child	**patient**
3	new	car	water	send	sell	**city**	price	year	request	list
4	use	can	get	chip	output	one	input	work	need	line
5	**include**	**offer**	price	**sale**	use	new	dos	good	**ship**	**sell**
6	appear	cover	st	copy	new	**issue**	th	**book**	vs	man
7	get	can	good	problem	time	**cause**	take	**doctor**	one	say
8	**wire**	use	**grind**	**current**	power	can	good	circuit	one	**amp**
9	**car**	good	**drive**	get	use	**engine**	**speed**	much	time	just
10	**know**	get	can	like	just	**anyone**	may	**thank**	**please**	make
11	good	year	**game**	**team**	run	**player**	get	**hit**	**win**	think

m = 10일 때 선택된 단어들은 Table 2에 굵은 글자체로 표현되어 있다. 로이터 기사 데이터에서는 “say”, “mln”, “company”, “dlrs” 등, 뉴스그룹 게시물 데이터에서는 “good”, “use”, “can”, “get” 등 다수의 토픽에서 공통적으로 관찰되는 단어들이 제거된 것을 확인할 수 있다. 즉, 단어 선택 절차를 거치면 많은 토픽에서 공통적으로 관찰되는 단어들이 제거되고 토픽 내용이 더 명확해지는 것을 알 수 있다. 예를 들어 뉴스그룹 게시물 데이터의 두 번째 토픽의 경우 “disease”, “medical”, “health”, “cancer”, “patient” 등의 단어가 선택되고 의학과 관련된 토픽이라는 것이 뚜렷해진다. 또, “include”, “offer”, “sale”, “ship”, “sell” 등이 선택된 다섯 번째 토픽의 경우 판매와 관련된 토픽, “car”, “drive”, “engine”, “speed” 등이 선택된 아홉 번째 토픽의 경우 자동차와 관련된 토픽이라는 것을 확인할 수 있다.

3.2절에서 제안한 절차1의 과정4와 같이 각 토픽별로 상위 10개씩의 β값과 카이제곱 통계량에 해당되는 단어들의 교집합으로 단어들이 선택되는 과정은 Figure 2를 통해 확인할 수 있다. 그림에서 β값과 카이제곱

Table 3: Number of documents and number of allocated topics (n)

data	n	q								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Reuters	0	0	1	90	433	891	1362	1742	2119	2429
	1	885	1603	2466	2603	2257	1786	1406	1029	719
	2	1003	1279	588	112	0	0	0	0	0
	3	827	259	4	0	0	0	0	0	0
	4	364	6	0	0	0	0	0	0	0
	5	64	0	0	0	0	0	0	0	0
	6	5	0	0	0	0	0	0	0	0
20 News Groups	0	0	1	45	332	816	1264	1608	1861	2035
	1	239	680	1405	1690	1382	934	590	337	163
	2	892	1212	740	176	0	0	0	0	0
	3	779	297	8	0	0	0	0	0	0
	4	247	8	0	0	0	0	0	0	0
	5	41	0	0	0	0	0	0	0	0

통계량이 동시에 큰 단어들이 선택되고 두 값 중 하나가 상대적으로 작은 단어들은 선택되지 않는 것을 알 수 있다. 예를 들어 Figure 2 (a)는 로이터 데이터의 두 번째 토픽에 해당되는 단어 선택 과정으로 LDA 모형 토픽-단어 행렬의 상위 10개 단어 중 카이제곱 통계량 값이 큰 “foreign”, “government”, “country”, “export”, “Brazil” 등의 단어는 선택되고 카이제곱 통계량 값이 상대적으로 작은 “say”, “debt”, “year”, “world”, “budget” 등의 단어는 선택되지 않은 것을 볼 수 있다.

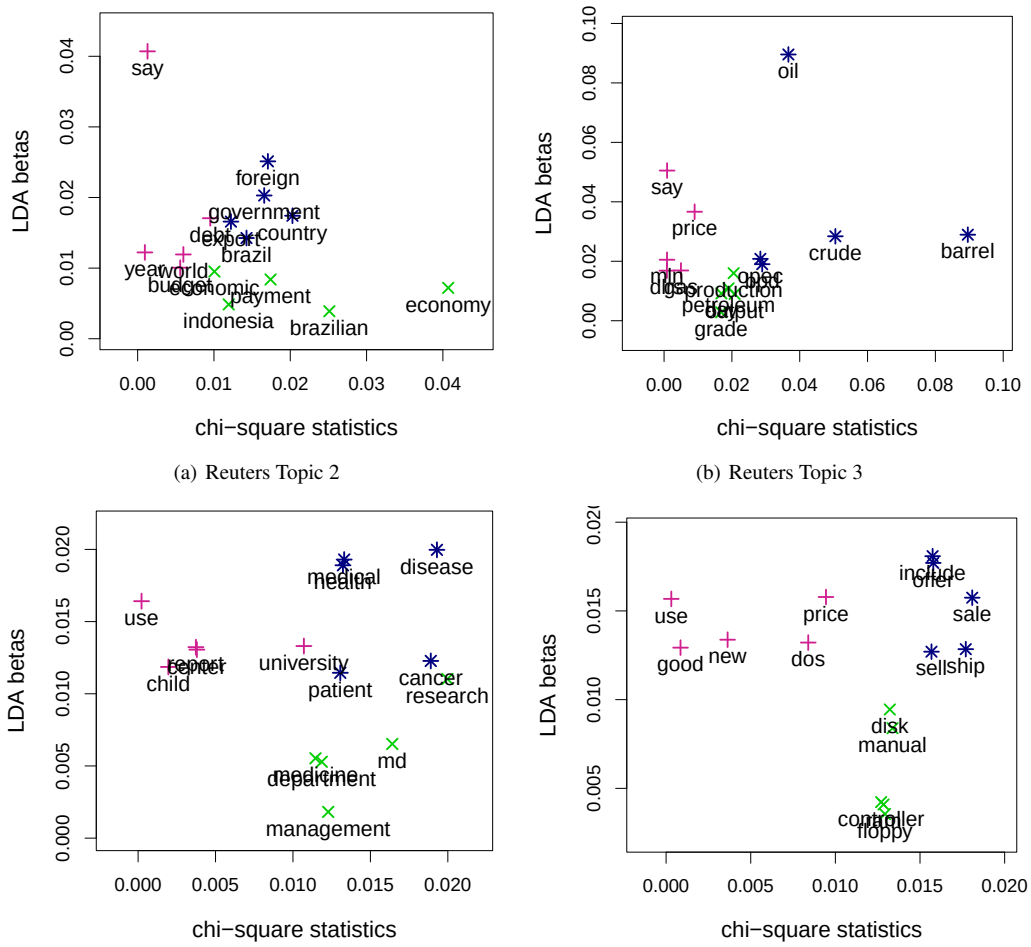
4.2. 선택된 단어를 이용한 군집 분석

이와 같은 절차를 통해 선택된 단어들을 이용하여 텍스트 데이터의 군집 분석을 실시해본다. 군집 분석을 위한 단어 선택 방법으로는 다음 두 가지 방법을 고려하였다.

- LDA-beta(ℓ) : 범주에 대한 정보를 사용하지 않고 LDA만을 이용하여 토픽별 상위 출현 빈도 단어 ℓ개씩을 선택
- LDA-beta-FS(ℓ) : 범주에 대한 정보를 사용하지 않고 LDA를 이용하여 구한 토픽별 ℓ개씩의 상위 출현 빈도 단어들 중 카이제곱 통계량 값이 큰 ℓ개 단어에 해당하는 단어들을 선택

예를 들어 Table 2에 기록된 모든 단어들은 LDA-beta(10)에 해당되는 단어들이고 Table 2에서 굵은 글자체로 표현된 단어들은 LDA-beta-FS(10)에 해당되는 단어들이다. 한편, 위의 정의에서 LDA-beta-FS(ℓ)에 해당되는 단어 집합은 LDA-beta(ℓ)에 해당되는 단어 집합의 부분 집합이 된다. 즉, LDA-beta-FS(ℓ)은 LDA-beta(ℓ)에 비해 항상 같거나 적은 수의 단어를 포함한다.

군집 분석 전에 먼저 문서-단어 행렬(DTM)을 단어빈도-역문서빈도(term frequency inverse-document frequency; TF-IDF) 형태의 행렬로 변환하고 변환된 행렬의 각 문서 사이의 거리를 코사인 비유사도(cosine dissimilarity) 기준으로 측정하였다. 군집 분석을 위해서는 K-중심값(K-medoids) 알고리즘을 이용하고 군집의 수는 여섯 개로 지정하였다. 범주의 수가 다섯 개이지만 군집의 수를 여섯 개로 지정한 이유는 Table 4에서 확인할 수 있는 것과 같이 군집의 수가 5일 때보다 6 이상일 때 군집 분석 결과와 목표 변수인 범주 사이의 연관성이 상대적으로 높게 나타났기 때문이다. 여기서 범주와 군집의 연관성은 각 군집 내의 다수 개체가 속하는 범주를 해당 군집의 범주로 간주하고 군집과 범주가 일치하는 개체의 비중으로 측정하였다. 군집의 수가 6보다 클 때 연관성이 소폭 높아지는 경우도 있지만 연관성이 높아지는 정도는 크지 않았다. 단어 선택



(a) Reuters Topic 2 (b) Reuters Topic 3

(c) 20 News Groups Topic 2 (d) 20 News Groups Topic 5

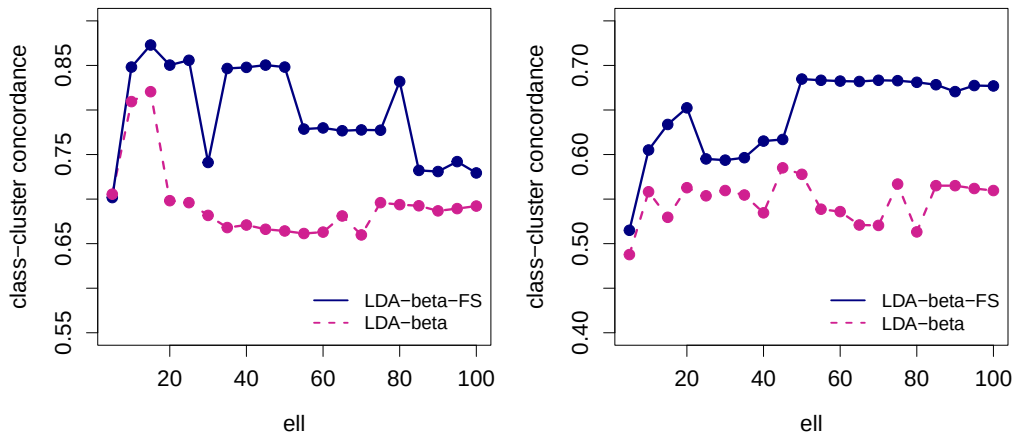
* : selected words, + : not selected top-10 β valued words, x : top-10 chi-squares valued words

Figure 2: *Word selection.*

Table 4: Class-cluster concordance for various number of clusters

# of clusters	LDA-beta						LDA-beta-FS					
	5	6	7	8	9	10	5	6	7	8	9	10
Reuters	0.805	0.821	0.814	0.811	0.812	0.838	0.838	0.873	0.865	0.845	0.845	0.850
News Groups	0.525	0.578	0.585	0.586	0.591	0.590	0.593	0.685	0.680	0.687	0.702	0.688

을 위한 LDA-beta-FS(ℓ)의 ℓ값도 여러 값들 중 가장 좋은 범주-군집 연관성을 나타내는 값으로 선택하였다. Figure 3에서 확인할 수 있는 것과 같이 로이터 기사 데이터의 경우 LDA-beta-FS와 LDA-beta 모두 $\ell = 15$일 때 범주와 토픽의 연관성이 가장 높고, 뉴스그룹 게시물 데이터의 경우 LDA-beta-FS는 $\ell = 50$, LDA-beta는 $\ell = 45$일 때 범주와 토픽의 연관성이 높은데 LDA-beta의 $\ell = 45$와 $\ell = 50$에서의 연관성 차이가 크지 않기 때문에 $\ell = 50$으로 결정하였다.



(a) Reuters data (b) News groups data

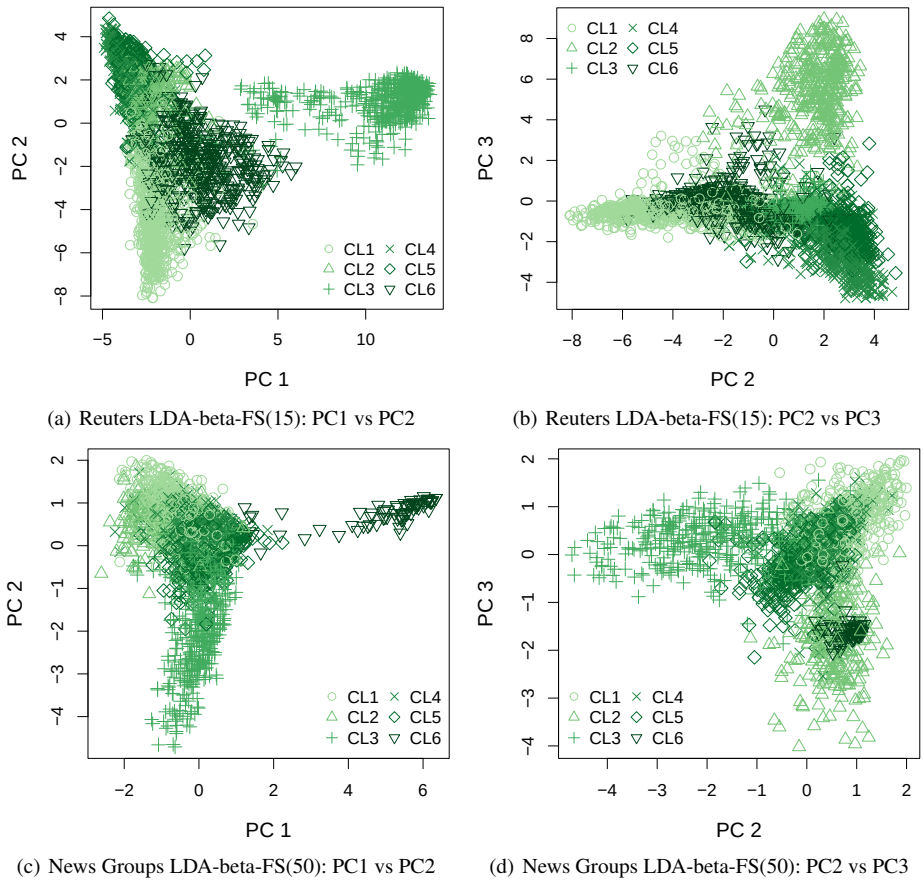
Figure 3: *Class-cluster concordance for various ℓ values.*

Table 5: Clustering results with 6 clusters

Cluster index	<Reuters Data>											
	LDA-beta(15)						LDA-beta-FS(15)					
	1	2	3	4	5	6	1	2	3	4	5	6
acq	490	20	5	24	19	219	610	11	2	32	21	101
crude	7	393	0	12	19	23	6	417	0	5	15	11
earn	45	61	566	5	2	326	70	11	535	15	0	374
money-fx	5	9	0	423	49	4	7	4	0	435	40	4
trade	0	11	2	24	385	0	0	5	0	30	377	10
Cluster index	<20 News Groups Data>											
	LDA-beta(50)						LDA-beta-FS(50)					
	1	2	3	4	5	6	1	2	3	4	5	6
misc.forsale	204	58	56	92	35	0	291	18	30	73	31	2
rec.autos	19	275	60	29	40	0	45	224	23	65	62	4
rec.sport.baseball	6	60	286	10	24	0	9	3	330	12	32	0
sci.electronics	52	93	61	185	65	0	96	13	20	268	58	1
sci.med	6	80	58	24	257	63	7	7	39	43	323	69

Figure 4에는 로이터 기사 데이터에 대해 LDA-beta-FS(15)로 선택한 단어들을 이용하여 군집 분석한 결과와 뉴스그룹 게시물 데이터에 대해 LDA-beta-FS(50)으로 선택한 단어들을 이용하여 군집 분석한 결과가 제시되어 있다. 군집 분석 결과를 그림으로 나타내기 위해 변환된 DTM을 주성분 분석을 통해 분해하고 각 주성분을 하나의 차원으로 하는 산점도에 각 군집의 점들을 표현하였다. 두 데이터 모두 산점도에서 각 군집들에 포함된 점들이 집단을 이루고 있는 것을 확인할 수 있으므로 군집 분석이 상당히 잘 이루어졌다는 것을 알 수 있다.

군집 분석 결과를 보다 객관적으로 판단하기 위해 군집 분석에 의해 결정된 군집과 각 데이터의 범주 사이에 연관성이 있는지 세부 범주와 군집별로 확인해본다. 군집과 범주 사이의 관계를 정리한 결과가 Table 5에 기록되어 있다. 표에서 확인할 수 있듯이 목표 변수에 대한 정보를 이용하지 않은 군집 분석에서 생성된 군집



(a) Reuters LDA-beta-FS(15): PC1 vs PC2

(b) Reuters LDA-beta-FS(15): PC2 vs PC3

(c) News Groups LDA-beta-FS(50): PC1 vs PC2

(d) News Groups LDA-beta-FS(50): PC2 vs PC3

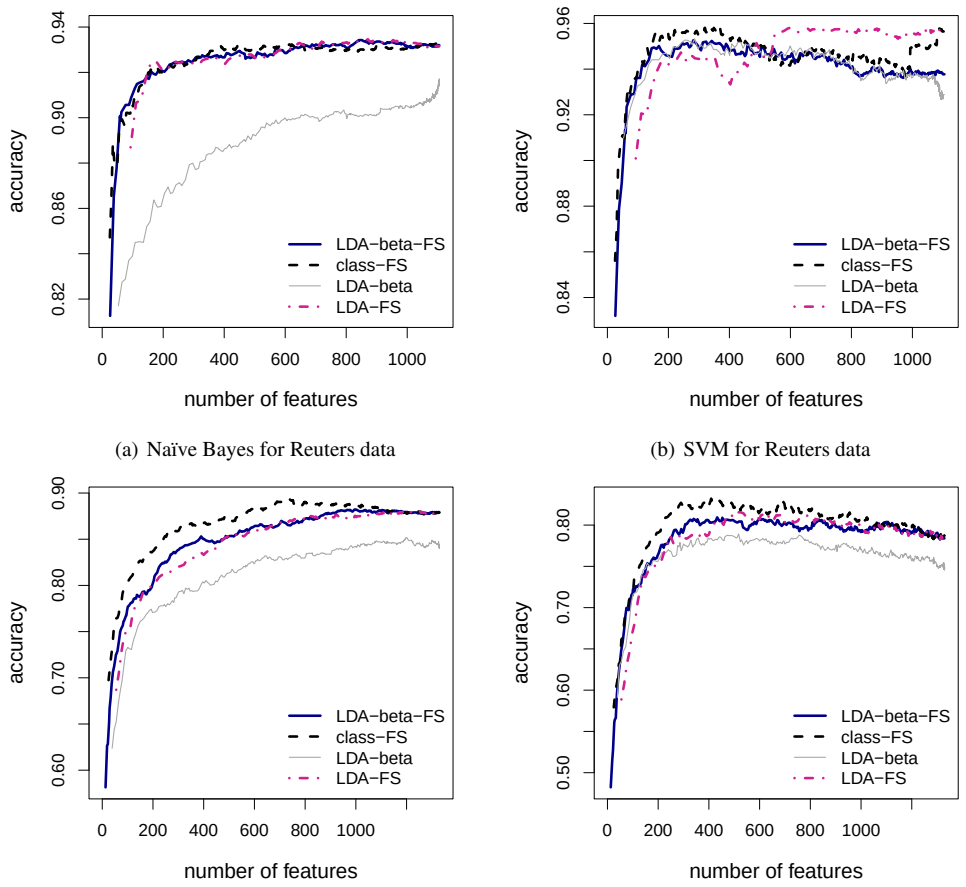
Figure 4: *Clustering results with 6 clusters.*

들이 목표 변수인 범주 값을 상당히 잘 반영하고 있음을 알 수 있다. 특히 LDA-beta를 이용한 군집 분석 결과보다 LDA-beta-FS를 이용한 군집 분석에서 군집과 범주 사이의 연관성이 높게 나타나는 것을 확인할 수 있다. 로이터 기사 데이터에서는 전체 3,148건의 기사 중 군집과 범주가 일치하는 기사가 LDA-beta-FS(15)의 경우 2,748건, LDA-beta(15)의 경우 전체 2,538건으로 범주와 군집의 일치도가 각각 87.3%, 82.1%로 나타났다. 또, 뉴스그룹 게시물 데이터에서는 전체 2,198건의 기사 중 군집과 범주가 일치하는 기사가 LDA-beta-FS(50)의 경우 1,505건, LDA-beta(50)의 경우 전체 1,270건으로 범주와 군집의 일치도가 각각 68.5%, 57.8%로 로이터 기사 데이터에 비해서 낮게 나타났다.

4.3. 선택된 단어를 이용한 분류 분석

3절에서 제안한 단어 선택 방법이 분류 분석에서도 잘 적용되는지 확인하기 위해 분류 분석을 실시해본다. 분류 분석을 위한 단어 선택 방법으로는 위의 두 가지 방법과 함께 아래와 같은 방식의 단어 선택 방법을 추가로 고려한다.

- LDA-FS(ℓ) : 범주에 대한 정보를 사용하지 않고 LDA만을 이용하여 토픽을 생성한 후 해당 토픽과 상관관계가 높은 ℓ개의 단어들을 카이제곱 통계량을 이용하여 선택



(a) Naïve Bayes for Reuters data

(b) SVM for Reuters data

(c) Naïve Bayes for 20 news groups data

(d) SVM for 20 news groups data

Figure 5: *Classification results for top-ℓ features.*

- class-FS(ℓ) : 분류 분석의 목표 변수인 범주에 대한 정보를 사용하여 각 범주별로 목표 변수와 높은 연관성을 가지는 ℓ개의 단어들을 카이제곱 통계량을 이용하여 선택

추가된 단어 선택 방법을 포함한 네 가지 단어 선택 방법 중 4.2절에서 고려한 LDA-beta-FS 등 다른 방법들은 분류 분석의 목표 변수인 범주에 대한 정보를 활용하지 않고 비지도 학습인 LDA 토픽 모형만을 이용하여 유용한 단어들을 선택하는 반면, class-FS의 경우 목표 변수에 대한 정보를 활용하고 있으므로 다른 세 방법에 비해 분류 분석에 있어서 더 좋은 분류 결과를 보일 것으로 기대할 수 있다.

이렇게 네 가지 방법으로 선택된 단어들을 나이브 베이즈 분류기와 지지 벡터 기계(support vector machine; SVM)에 적용하여 분류 분석을 실시해본다. SVM의 경우 오분류에 대한 비용 C와 커널(kernel)을 지정할 필요가 있는데 여러 값들 중 가장 좋은 결과를 보인 $C = 0.01$과 선형 경계를 가정하였다.

Figure 5에는 각 데이터에 분류 모형을 적용한 결과가 제시되어 있다. 먼저 로이터 기사 데이터에 나이브 베이즈 분류기와 SVM을 적용시킨 결과를 보면 LDA-beta-FS가 범주 정보를 이용한 class-FS와 거의 비슷한 정확도(accuracy)를 나타내는 것을 알 수 있다. 반면에 카이제곱 통계량을 이용한 단어 선택 과정 없이 토픽 모형만으로 단어를 선택한 LDA-beta의 경우 다른 방법들과 동일한 개수의 단어를 선택했을 때 나이브 베이

즈 분류 모형에서 정확도가 더 낮게 나타나는 것을 확인할 수 있다. LDA-beta-FS를 LDA-FS와 비교해보면 두 방법이 전반적으로 비슷한 정확도를 보였지만 SVM을 적용한 경우 선택된 단어의 개수 ℓ이 작을 때에는 LDA-beta-FS가 상대적으로 정확도가 높게 나타났다. 한편, 뉴스그룹 게시물 데이터에서는 LDA-beta-FS의 정확도가 class-FS보다 낮지만 큰 차이는 없고 LDA-beta에 비해서는 더 높은 정확도를 보이는 것을 알 수 있다. LDA-beta-FS와 LDA-FS는 전반적으로 비슷한 정확도를 보이지만 로이터 데이터에서와 같이 선택된 단어의 개수 ℓ이 작을 때 LDA-beta-FS의 정확도가 상대적으로 높은 경향이 있다.

정확한 분류를 위해 필요한 단어의 개수 측면에서는 데이터별로 다소 차이가 있다. 예를 들어 로이터 기사 데이터에서는 약 100개 이내의 단어만으로도 1,000개 이상의 단어들을 사용하여 분류할 때와 큰 차이가 없는 분류 결과를 얻을 수 있었다. 즉, 로이터 기사 데이터의 경우 3절에서 제안된 바와 같이 LDA-beta-FS 방식의 단어 선택 절차를 통해 적은 단어로 텍스트 데이터의 정보를 요약할 수 있음을 알 수 있다. 반면, 뉴스그룹 게시물 데이터에서는 1,000개 정도의 단어를 사용하였을 때와 비슷한 정확도를 얻기 위해서는 약 400개 이상의 상당히 많은 단어를 선택하여야 하는 것으로 나타났다.

5. 결론

이 연구에서는 목표 변수에 대한 정보가 없는 비지도 학습에서 단어 선택 방법에 대해 살펴보았다. 지도 학습에서는 목표 변수에 대한 정보를 이용하여 카이제곱 통계량 등의 지표를 구하여 단어를 선택하는 반면, 비지도 학습에서는 목표 변수에 대한 정보가 없으므로 토픽 모형을 통해 구한 토픽을 지도 학습에서의 목표 변수 대신 사용하여 단어를 선택하는 절차를 제안하였다. 이 연구에서 제안한 단어 선택 절차를 거치면 여러 토픽에 공통적으로 자주 등장하는 단어들이 제거되어 토픽을 더 명확하게 식별할 수 있다는 장점이 있다. 또, 목표 변수에 대한 정보 없이 토픽 모형을 이용하여 단어들을 선택하더라도 군집 분석 결과를 통해 구한 군집이 범주와 매우 높은 연관성을 가지고 있음을 확인하였다. 이와 더불어 분류 분석에 있어서도 목표 변수를 이용한 경우와 비슷한 정확성을 보이는 등 텍스트 데이터의 정보를 충분히 유지할 수 있음을 알 수 있었다.

References

Arun R, Suresh V, Madhavan CEV, and Murthy MN (2010). On finding the natural number of topics with latent Dirichlet allocation: Some observation, *Pacific-Asia Conference on Knowledge Discovery and Data Mining, Par I, LNAI*, **6118**, 391–402.

Blei DM, Ng AY, and Jordan MI (2003). Latent Dirichlet allocation, *Journal of Machine Learning Research*, **3**, 993–1022.

Blei DM (2012). Probabilistic topic models, *Communications of the ACM*, **55**, 77–84.

Boyd-Graber J, Hu Y, and Minmo D (2017). Applications of topic models, *Foundations and Trends in Information Retrieval*, **11**, 143–296.

Cao J, Xia T, Li J, Zhang Y, and Tang S (2009). A density-based method for adaptive LDA model selection, *Neurocomputing*, **72**, 1775–1781.

Deerwester S, Dumais ST, Furnas GW, Landauer TK, and Harshman R (1990). Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, **41**, 391–407.

Deveaud R, SanJuan E, and Bellot P (2014). Accurate and effective latent concept modeling for ad hoc information retrieval, *Document Numérique*, **17**, 61–84.

Griffiths TL and Steyvers M (2004). Finding scientific topics, *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 5228–5235.

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

Hofmann T (1999). Probabilistic latent semantic indexing, In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, California, 50–57.

Lee H, Choi J, Lee S, and Son W (2021). Topic change monitoring study based on Blue House national petition using a control chart, *The Korean Journal of Applied Statistics*, **34**, 795–806.

Mun HI and Son W (2022) Properties of chi-square statistic and information gain for feature selection of imbalanced text data, *The Korean Journal of Applied Statistics*, **35**, 469–484.

Son W (2020). Skewness of chi-square statistic for imbalanced text data, *Journal of the Korean Data and Information Science Society*, **31**, 807–821.

Received August 25, 2022; Revised September 28, 2022; Accepted October 10, 2022

토픽 모형을 이용한 텍스트 데이터의 단어 선택

장우솔[a], 김예은[a], 손원[1,a]

[a]단국대학교 대학원 응용통계학과

요 약

텍스트 데이터는 일반적으로 많은 변수를 포함하고 있으며 변수들 사이의 연관성도 높아 통계 분석의 정확성, 효율성 등에서 문제가 생길 수 있다. 이러한 문제점에 대처하기 위해 목표 변수가 주어진 지도 학습에서는 목표 변수를 잘 설명할 수 있는 단어들을 선택하여 이 단어들만 통계 분석에 이용하기도 한다. 반면, 비지도 학습에서는 목표 변수가 주어지지 않으므로 지도 학습에서와 같은 단어 선택 절차를 활용하기 어렵다. 이 연구에서는 토픽 모형을 이용하여 지도 학습에서의 목표 변수를 대신할 수 있는 토픽을 생성하고 각 토픽별로 연관성이 높은 단어들을 선택하는 단어 선택 절차를 제안한다. 제안된 절차를 실제 텍스트 데이터에 적용한 결과, 단어 선택 절차를 이용하면 많은 토픽에서 공통적으로 자주 등장하는 단어들을 제거함으로써 토픽을 더 명확하게 식별할 수 있었다. 또한, 군집 분석에 적용한 결과, 군집과 범주 사이에 높은 연관성을 가지는 군집 분석 결과를 얻을 수 있는 것으로 나타났다. 목표 변수에 대한 정보없이 토픽 모형을 이용하여 선택한 단어들을 분류 분석에 적용하였을 때 목표 변수를 이용하여 단어들을 선택한 경우와 비슷한 분류 정확성을 얻을 수 있음도 확인하였다.

주요용어: 변수 선택, 비지도 학습, 잠재 디리클레 할당, 텍스트 데이터, 토픽 모형

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