The Korean Journal of Applied Statistics
2023, Vol. 36, No. 3, 211–223

DOI: https://doi.org/10.5351/KJAS.2023.36.3.211
Print ISSN1225-066X / Online ISSN 2383-5818

Predicting movie ratings using Korean reviews with multinomial inverse regression

Hyunjin Kim[a], Ji Su Kim[a], Eun Ryung Lee[1,a]

[a]Department of Statistics, Sungkyunkwan University

Abstract

With the development of IT technology, a vast amount of information, including articles, blogs and online comments, is being circulated in modern society. Much of this data consists of text, which is unstructured data and is constantly being produced. As a result, the demand for analysis using text mining continues to increase by converting unstructured data into new digitized data. Unlike data mining, which extracts patterns from the structured data, the biggest difference is that text mining looks for patterns from the unstructured text. A main purpose of this study is to examine how high-dimensional Korean text variables, e.g., Korean movie reviews, can be summarized to provide sufficient information and to reduce dimension. This helps a further analysis, e.g., prediction of movie ratings. We use a state of the art method called multinomial inverse regression (MNIR), which was proposed to obtain a low-dimensional summary of texts that contains sentiment information. This requires the review comments to convert into an appropriate form for the model, so it was preprocessed and quantified in various ways. Besides prediction evaluation resulting from the summarized information, we propose the multiple MNIR models that consider interaction terms with movie genre variables. The proposed methods reveal an interesting feature of text data and it show a better prediction performance in our empirical study.

Keywords: text mining, sufficient dimension reduction, marginal models, multinomial inverse regression, multiple responses, sentiment analysis, review rating prediction

1. 서론

방대한 양의 자료와 온라인상의 댓글은 현대 IT 기술이 발달되면서 기하급수적으로 끊임없이 증가하고 있으며, 기계학습 방식으로 텍스트를 분석하는 일은 그 자체만으로 많고 방대한 내용을 채울 수 있다. 정보 검색, 데이터 마이닝, 기계 학습, 통계학, 그리고 컴퓨터 언어학 등 여러 분야의 방법을 이용하여 흥미롭고 새로운 정보를 추출하는 방법 (Gupta와 Lehal, 2009)인 여러 텍스트 마이닝 기법이 소개되었지만, 텍스트 마이닝에서 영어와 달리 한국어의 경우 텍스트 처리 기술에 대한 연구는 상대적으로 부족한 상태이다. 용언의 불규칙한 사용, 자유로운 어순, 맞춤법 파괴 등과 같은 한글 특성으로 인해 식별할 수 있는 댓글을 알아내는 데 어려움이 존재하기 때문이다. 본 연구에서는 예측변수가 한국어 텍스트일 때, 이 예측변수들의 정보가 포함된 저차원의 점수로도 감성(sentiment) 변수를 예측할 수 있는지 알아보고자 한다. 또한, 한국어 텍스트를 정밀하게 전처리하여 분석하는 것이 중요하기 때문에 전처리 과정에 대해서도 상세히 서술하고자 한다.

This work is supported by a National Research Foundation of Korea grant funded by the Korean government (no. NRF-2022R1A2C1012798).

[1] Corresponding Author: Department of Statistics, Sungkyunkwan University, 25-2 Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: silverryuee@gmail.com

Published 30 June 2023 / journal homepage: http://kjas.or.kr

본 연구에 사용된 자료는 평점이 있는 영화 데이터이다. 영화 데이터를 통해 감성 분석을 수행할 예정이며, 구체적으로 다항 역회귀(multinomial inverse regression) (Taddy, 2013) 기법을 활용하여 댓글의 요약된 정보인 저차원의 점수를 얻고 이를 활용하여 영화평점이라는 감성변수를 예측하고자 한다. 더불어 기존 다항 역회귀에서 생기는 한계점을 파악하고, 이를 해결하기 위한 방법으로 장르라는 외생변수(exogeneous variable)를 추가적으로 고려하여 교차항이 포함된 다중 반응변수(multiple response variables)를 생성해 여러 다항 역회귀 모형들을 고려하는 것을 제안한다. 이렇게 기존 모형과의 비교 분석을 실시하여 향상된 점을 기술하고, 분석 결과를 통합하여 다른 분야에 추가로 활용 및 응용될 수 있는 방향을 제시하고자 한다.

본 연구에서 활용한 다항 역회귀는 역회귀(inverse regression) 기법의 일종으로, 충분 차원 축소라는 개념이 사용된다. 먼저 차원 축소(dimension reduction)는 변수의 차원이나 자료를 특정 목적에 따라 줄이는 방법으로서, 모형의 복잡성(complexity)을 감소시키기 위해 사용된다. 차원 축소를 통해 관측 자료를 잘 설명하는 잠재 공간(latent space)을 찾을 수 있으며, 대표적인 방법으로는 주성분 분석(principal components analysis; PCA)과 특이값 분해(singular value decomposition; SVD)가 있다. 충분차원축소(sufficient dimension reduction; SDR)는 통계학에서 차원 축소와 충분성(sufficiency)의 개념을 합하여 자료를 분석하는 형태이다.

일반적으로 충분차원축소는 역회귀를 사용하며, 대표적인 SDR 방법들 중에는 분할 역회귀(sliced inverse regression; SIR) (Li, 1991) 와 SAVE (sliced average variance estimation) (Cook과 Weisberg, 1991)가 있다. 텍스트 자료를 독립변수 $\mathbf{x}$로 활용하여 감성변수(sentiment variable) y를 예측하는 것은 $\mathbf{x}$의 초고차원 성질 때문에 통계적으로 매우 어려운 문제이다. 이러한 어려움을 극복하기 위해 충분차원축소를 활용한 다항 역회귀 모형이 최근 제안되었다 (Taddy, 2013).

본 논문은 총 5장으로 다음과 같다. 먼저 제 2장에서는 영화 평점자료를 소개하고, 전처리 과정을 서술한다. 제 3장에서는 분석을 수행하기 위한 선행 연구 내용을 소개하고, 교차항이 포함된 추가 모형들을 제안한다. 제 4장에서는 전처리된 영화 자료를 분석한 결과를 설명하고, 본 연구에서 제안한 방법을 기존의 방법과 비교한다. 마지막으로는 본 연구에 대한 결론과 논의점을 언급하며 마무리한다.

2. 분석 자료 및 전처리

2.1. 영화 평점 자료

본 연구에서 분석한 자료는 140자 이내의 네이버 영화평 데이터로 온라인 상에서 (`github.com/drexly/movie140reviewcorpus`) 얻을 수 있다. 주어진 자료에서 변수는 영화평 식별 아이디, 평점, 영화평 추천과 비추천 수, 그리고 네이버 영화평으로 이루어져 있다. 기존 변수 외에 제목에 쓰여져 있는 영화 장르와 영화 식별 아이디를 가져와 새로운 변수로 추가하였다. 분석에 사용된 변수는 장르, 영화 식별 아이디, 1점부터 10점으로 이루어져 있는 평점, 그리고 영화평 텍스트이다. 그 결과 불러온 자료의 관측치는 총 9,541,227개이며, 장르 종류의 개수는 1,583개이다. 영화 평점 자료를 불러오는 과정에서 사용된 4개의 변수에서 결측치가 있는 경우 분석 대상에서 제외한 이후 분석 모형에 적합하기 위한 훈련 자료를 얻기 위해 1% 표본 추출을 하였고, 추출된 관측치에서 2.2절의 전처리를 수행한 후 16,973개의 관측치가 분석에 사용되었다. 또한 10점인 평점이 압도적으로 많은 불균형한 데이터(imbalanced data)임을 고려하여 1점부터 3점은 1점(부정), 4점부터 7점은 2점(중립), 8점부터 10점은 3점(긍정)인 3개의 감성 카테고리로 변환하였다. 변환 후 부정, 중립, 긍정의 감성을 나타내는 관측치의 개수는 각각 2,767개, 2,947개, 그리고 11,259개이다.

2.2. 영화평 텍스트 전처리 과정

일반적으로 텍스트 자료 분석에서 통계적으로 더 정확하고 의미있는 결과를 얻기 위해서 텍스트 자료를 잘 정제하는 것이 중요하며, 이는 상당한 전처리 과정을 요구한다. 영어로 이루어져 있는 텍스트의 경우, 어간

(stemming word)추출과 같은 전처리 방법들이 어느 정도 자리를 잡았지만, 한국어의 경우 언어학적으로 영어와 다른 구조을 가지며 훨씬 더 복잡하여 자연어 처리가 어렵다는 것이 알려져 있다. 전처리 과정으로 띄어쓰기, 형태소 분석 및 사전 구축을 수행하는 R 상에 있는 한국어 자연어 처리 알고리즘들을 활용했고 추가적으로 사전에 등록되지 않는 단어를 추가하는 등 작업을 직접 수행하여 보다 정제된 영화평 텍스트 자료를 얻고자 했으며, 자세한 방법은 다음과 같다. 한국어는 띄어쓰기가 제대로 되어있지 않아도 의미 전달이 쉽다는 특성이 있다. 특히 온라인 상에서는 띄어쓰기가 잘 되어있는 경우를 찾기 어려운데 이는 텍스트 자료 전처리 과정인 형태소 분석 및 사전 구축을 어렵게 만든다. 본 연구에서는 자동 띄어쓰기를 해주는 딥러닝 모델 R 패키지인 KoSpacing (Jeon, 2018)을 사용하였다. 또한, R의 한글 자연어 분석 패키지인 KoNLP (Korean natural language processing) (Jeon, 2016)를 활용하여 형태소 분석을 실시하고, 태깅된(tagging) 품사의 단어를 등록된 사전을 통해 추출하였다. 어절을 분석하여 명사만 품사를 붙여주는 데 그치지 않고 한나눔에서 기본적으로 사용하는 카이스트 형태소 태그 집합인 KAIST 품사 태그셋을 활용하여 보통명사(NC), 형용사(PA), 동사(PV)를 추가로 추출하였다.

KoNLP 패키지 내의 국립국어원에서 제공하는 세종(Sejong) 사전(8만개)과 추가로 한국정보화진흥원에서 제공하는 형태소 사전 NIA사전(93만개)을 활용하여 한국어 표현 분해 및 단어 분류를 할 수 있다. 이 외에 사전에 등록되지 않은 단어들을 직접 파악하고 사전에 추가하여 함께 분석에 사용하였다. 형태소 분석에는 'SimplePos22' 함수를 사용하여 품사를 확정시키고 각 단어들의 태그를 붙여주었다. 그중에서 동사의 경우, '-다'가 빠진 형태로 표시되기 때문에 동사인 단어들은 동사 형태로 표현하기 위해 다시 '-다'를 추가하였다. 기본적으로 두 글자 미만의 길이인 단어는 분석에서 제외하였으나, 단어 길이가 한 글자이지만 '굿', '잼', '짱'과 같은 용어처럼 빈도수가 100 이상이고 충분히 의미가 있다고 판단되는 단어는 직접 따로 추출하여 포함시켰다. 마찬가지로 형태소 분석을 한 뒤 영화평 내용이 없는 경우는 제거하였다. 이후 한국어가 아닌 다른 언어인 영어, 중국어, 일본어 등을 제거하고 분석에 의미가 없는 리뷰 내의 구두점, 숫자, 특수문자, 한글 불용어(stopwords) 역시 사전에서 제외하였다. 또한 'ㅋㅋ'와 'ㅠㅠ' 같이 단순히 자음 혹은 모음만 있는 경우 역시 완벽한 형태가 아니므로 이 단어들 또한 삭제하였다. 댓글 내에서 줄 바꿈이 있는 경우 프로그램 상에서 특정 기호가 출력되기 때문에 직접 제거했고, 이렇게 제거했을 시 앞뒤와 글자 사이사이에 공백이 생기는 등 2개 이상의 공백이 있는 경우가 있어 처리하였다. 마지막으로, 텍스트 마이닝의 가장 대표적인 패키지인 tm 패키지 (Feinerer, 2017)를 사용하여 텍스트 자료를 말뭉치로 변형하고, 정제된 말뭉치로부터 단어-문서 행렬을 생성하여 분석에 활용했다.

3. 텍스트 자료 분석 모형 및 방법

3.1. 텍스트 자료 표현: 단어-문서 행렬

텍스트 자료는 종종 단어-문서 행렬(term document matrix; TDM)이라는 행렬 형태로 변환되어 분석된다. 이 TDM 행렬에서 각 행과 열은 문서(documents)와 단어(term)를 의미하고, (i, j) 원소는 문서 i에서 단어 j가 발생하는 빈도수를 나타내게 된다. 예를 들어, Table 1에는 나와 있는 한 문장으로 구성된 3개의 문서를 고려해보자. 문서 1을 자세히 보면 5개의 단어로 구성되어 있으며 다음과 같은 빈도 벡터

$$\boldsymbol{x}_1 = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 2 \end{bmatrix}^T \tag{3.1}$$

Table 1: Example of documents

No.	Documents
Doc 1	the fox chased the rabbit
Doc 2	the rabbit ate the cabbage
Doc 3	the fox caught the rabbit

Table 2: Document term matrix representation of example documents

	Ate	Cabbage	Caught	Chased	Fox	Rabbit	The
Doc 1 ($\mathbf{x}_1$)	0	0	0	1	1	1	2
Doc 2 ($\mathbf{x}_2$)	1	1	0	0	0	1	2
Doc 3 ($\mathbf{x}_3$)	0	0	1	0	1	1	2

로 표현 될 수 있다. 따라서, Table 1에 나와 있는 세개의 문서는 Table 2로 정리할 수 있고, 다음과 같은 TDM 행렬

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 2 \\ 1 & 1 & 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 1 & 0 & 1 & 1 & 2 \end{bmatrix} \tag{3.2}$$

로 요약된다. 일반적으로 텍스트 분석에서 사용되는 문서들은 수많은 문장들을 포함하고 분석에 사용되는 총 단어의 수는 수천 혹은 수만개로 TDM 행렬의 차원은 매우 크다. 연구에서 분석한 네이버 영화평 텍스트 자료는 2장에서 소개한 전처리 과정을 거친 후, 18,920개의 단어와 16,973개의 문서(영화평)로 이루어지며 모든 단어들의 총 사용빈도는 71,385이다.

3.2. 다항 역회귀 모형(multinomial inverse regression; MNIR)

문서 i(예를 들어, 영화 평점 자료에서 아이디i의 영화평)의 빈도벡터와 감성변수(예를 들어, 영화 평점 자료에서 평점)를 각각, $\mathbf{x}_i$, y_i라고 두고, 총 빈도는 $m_i = \sum_j x_{ij}$ 라고 놓겠다. Taddy (2013)에서 고려한 다항 역회귀 모형은

$$\mathbf{x}_i \mid m_i, y_i \sim \text{MN}(m_i, q_i) \quad \text{with} \quad q_{ij} = \frac{\exp\left(\eta_{ij}\right)}{\sum_{i=1}^{J} \exp\left(\eta_{ij}\right)} \tag{3.3}$$

을 가정한다. 다시 말해, 총 빈도 m_i와 감성변수 y_i가 주어졌을 때 빈도 벡터 $\mathbf{x}_i$는 크기 m_i와 확률 $q_i = (q_{ij} : j = 1, \ldots, J)^\top$를 가지는 J-차원 다항분포를 따른다. 여기에서, $\eta_{ij} = \alpha_j + \mathbf{u}_{ij} + \mathbf{v}_i^T \varphi_j$로 주어지며, α_j는 절편항, $\mathbf{u}_i = (\mathbf{u}_{ij} : j = 1, \ldots, J)^\top$는 임의효과(random effect), $\mathbf{v}_i$는 반응변수와 관련 있는 K-차원 반응 인자(response factor), $\Phi = [\varphi_1, \ldots, \varphi_K]$는 역회귀 계수의 $p \times K (K \ll p)$ 행렬을 나타내며 $\mathbf{v}_i$는 조건부 분포 $\boldsymbol{x}_i | y_i$에 대한 충분성을 가정한다, 즉, $y_i \perp\!\!\!\perp \mathbf{x}_i | v_i$ 이다. Taddy (2013)은 모형 (3.3) 하에서 상대 빈도(relative frequency) $\boldsymbol{f}_i = \mathbf{x}_i / m_i$를 활용한 충분축소 점수(sufficient reduction score) $\Phi^\top \boldsymbol{f}_i$를 정의할 수 있고 이는 $y_i \perp\!\!\!\perp \mathbf{x}_i | \Phi^\top \boldsymbol{f}_i$가 성립한다는 것을 보였다. 다시 말해, 모형 (3.3) 하에서 $y|\mathbf{f}, m$의 조건부 분포는 $y|\Phi^T\mathbf{f}, m$의 조건부 분포와 같으므로, 초고차원 상대 빈도벡터 $\mathbf{f}_i$ 대신 차원 축소된 (추정된) 충분축소점수 $\hat{\Phi}^T\mathbf{f}_i$를 활용하여 평점 y을 예측할 수 있다. 실제 영화평점 자료 분석에서 영화평 텍스트의 충분축소 점수를 활용하여 영화평점을 예측했으며, 이는 4.3절에 자세히 기술되어 있다.

충분축소 점수를 형성하는 Φ의 추정량은 반응인자 $\mathbf{v}_i$에 따라 달라진다. Taddy (2013)이 제안한 다항 역회귀 방법은 y_i의 임의의 함수인 $\mathbf{v}_i$에 대해 일반적으로 적용 가능하지만, $\mathbf{v}_i$의 결정에 대해 엄밀한 논의를 하지

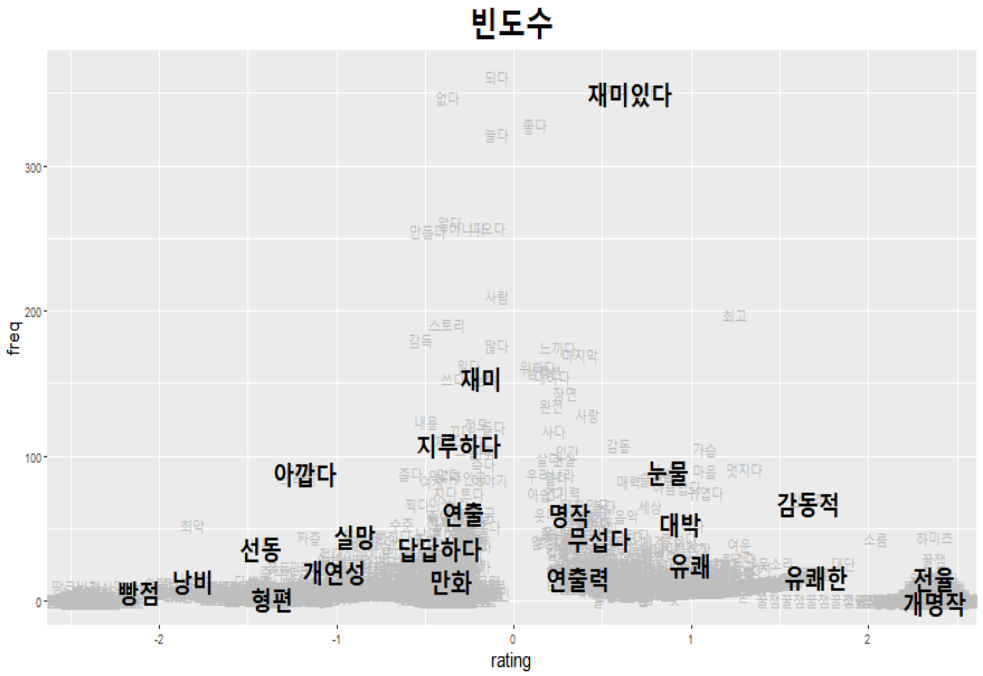

Figure 1: *Graph for term loadings in rating - frequency.*

않았다. 실제 분석에서 범주형 감성변수 y_i인 경우 $\mathbf{v}_i = y_i$를 잡았고, 연속형 감성변수인 경우 y_i의 계단함수로 $\mathbf{v}_i$를 제안했다. 이 외에도, 기존의 역회귀 방법들을 확장하여 y의 다항함수와 같은 다양한 함수 형태나 교차검증을 통해 $\mathbf{v}_i$를 결정하는 것을 고려할 수 있을 것이다 (Adragni와 Cook, 2009; Ma와 Zhu,2013; Tan 등 ,2018). 이에 대한 후속 분석을 고려하고 있다. $\hat{\Phi}$는 (3.3) 다항역회귀모형 (3.3)의 계수행렬 $\Phi^\top = (\phi_1, \ldots, \phi_J)$ 추정치로 R 패키지 'textir' 내의 함수 'mnlm'을 통해 계산할 수 있다. R 함수 'mnlm'은 절편항 α_j, 임의효과 $\mathbf{u}_{ij}$에 대한 다음과 같은 사전분포를, 그리고

$$\alpha_j \overset{i.i.d.}{\sim} N(0,1),$$
$$\exp\left(\mathbf{u}_{ij}\right) \overset{i.i.d.}{\sim} \text{Gamma}(1,1),$$

ϕ_j에 대한 라플라스 사전 분포(independent Laplace priors), 라플라스 분포의 비율모수(rate parameter) λ_j에 대한 감마 초사전분포(hyperprior)를 식 (3.4)와 같이 지정하여

$$\pi\left(\varphi_j, \lambda_j\right) = \frac{\lambda_j}{2} e^{-\lambda_j|\varphi_j|} \frac{r^s}{\Gamma(s)} \lambda_j^{s-1} e^{-r\lambda_j} \tag{3.4}$$

다항역회귀 (3.3)의 최대 사후분포추정치를 계산해준다. 라플라스 사전 분포는 ϕ_j에 대한 라쏘 벌점 방법(lasso penalty)으로 볼 수 있고 Φ에 대한 성긴 추정치를 주게 된다.

3.3. 제안: 장르에 의존하는 다항 역회귀 모형

장르에 따라 영화평 작성 단어의 사용이 달라지는 현상이 관측되어, 이에 따라 장르라는 외생변수에 의존하는 다항 역회귀 모형을 고려하고 이를 2절의 네이버 영화평 자료 분석에 사용하고자 한다. 구체적으로, 평점 R,

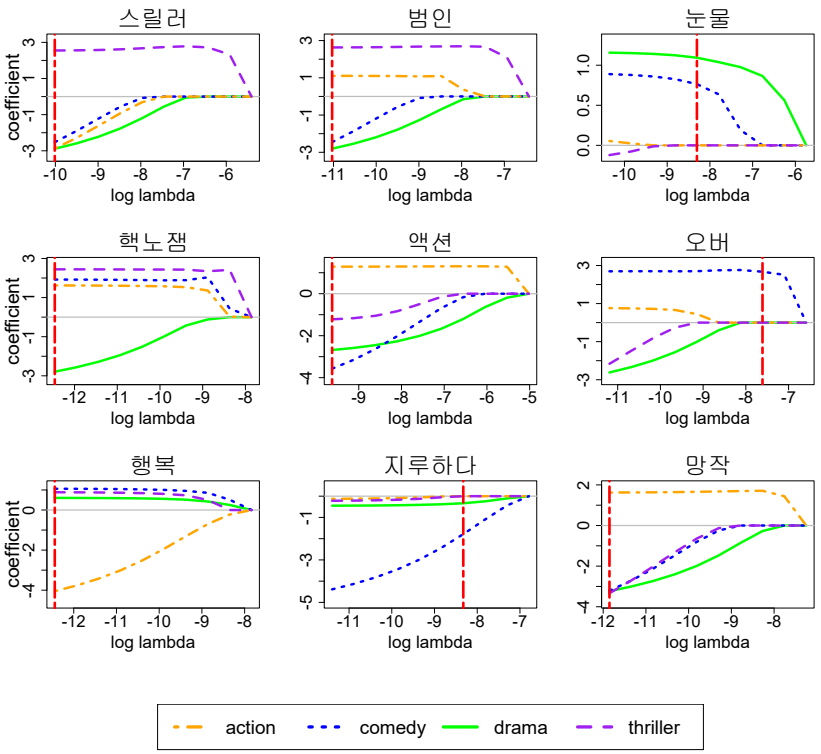

Figure 2: *Poisson regression regularization paths for words: Green, blue, orange, purple lines correspond to drama, comedy, action, thriller, respectively.*

장르 G 변수들에 대한 독립적인 주효과(main effect)뿐만 아니라 장르와 평점 변수의 상호작용 효과 까지 고려했고, 그 결과 생성된 다중 변수들 각각에 다항 역회귀 기법을 적용한다. 다음과 같은 다항 모형들을 가정한다.

$$\mathbf{x} \mid m, R \sim \mathrm{MN}\left(m, \mathbf{q}\left(v_R\right)\right) \quad \text{with} \quad \eta_{R,j} = \alpha_{R,j} + u_{R,j} + v_R^T \phi_{R,j}. \tag{3.5}$$

$$\mathbf{x} \mid m, G_k \sim \mathrm{MN}\left(m, \mathbf{q}\left(v_{G_k}\right)\right), \quad k = 1,2,3,4, \quad \text{with} \quad \eta_{G_k,j} = \alpha_{G_k,j} + u_{G_k,j} + v_{G_k}^T \phi_{G_k,j}. \tag{3.6}$$

$$\mathbf{x} \mid m, R * G_k \sim \mathrm{MN}\left(m, \mathbf{q}\left(v_{R*G_k}\right)\right), \quad k = 1,2,3,4, \quad \text{with} \quad \eta_{R*G_k,j} = \alpha_{R*G_k,j} + u_{R*G_k,j} + v_{R*G_k}^T \phi_{R*G_k,j}, \tag{3.7}$$

여기에서 R은 평점, G_1은 드라마, G_2는 코미디, G_3는 액션, 그리고 G_4는 스릴러 장르를 뜻하며, $v_R = R, v_{G_k} (k = 1,2,3,4)$는 장르 G_k에 대한 가변수(dummy variable)로 잡았다. 평점과 장르의 교차항들은 $v_{R*G_1} = v_R \times v_{G_1}$, $v_{R*G_2} = v_R \times v_{G_2}$, $v_{R*G_3} = v_R \times v_{G_3}$, $v_{R*G_4} = v_R \times v_{G_4}$ 라고 정의한다. 식 (3.5), (3.6), (3.7) 의 모형들은 고려하고 있는 반응변수 $R, G_k, R * G_k (k = 1,2,3,4)$ 들의 주변 모형(marginal model)이며 변수들 사이에 의존성을 허용한다. 텍스트 자료와 같은 의존성이 있는 반응변수가 고차원 일 때 결합분포(joint distribution)를 모델링 하는 것은 일반적으로 어려운 문제이며, 이러한 문제를 우회하기 위해 주변분포 모델링을 활용할 수 있다 (Jentsch 등, 2021). R 패키지 'textir' 내의 함수 'mnlm'을 통해 다항 역회귀들을 각각 적합했고 그 결과 생성된 9개의 충분점수들 $\hat{\Phi}_R^\top \boldsymbol{f}_i$, $\hat{\Phi}_{G_k}^\top \boldsymbol{f}_i$, $\hat{\Phi}_{R*G_k}^\top \boldsymbol{f}_i$ $(k = 1,2,3,4)$ 을 활용하여 영화평점 예측을 수행할 것이며, 이는 4.3절에 자세히 기술되어 있다.

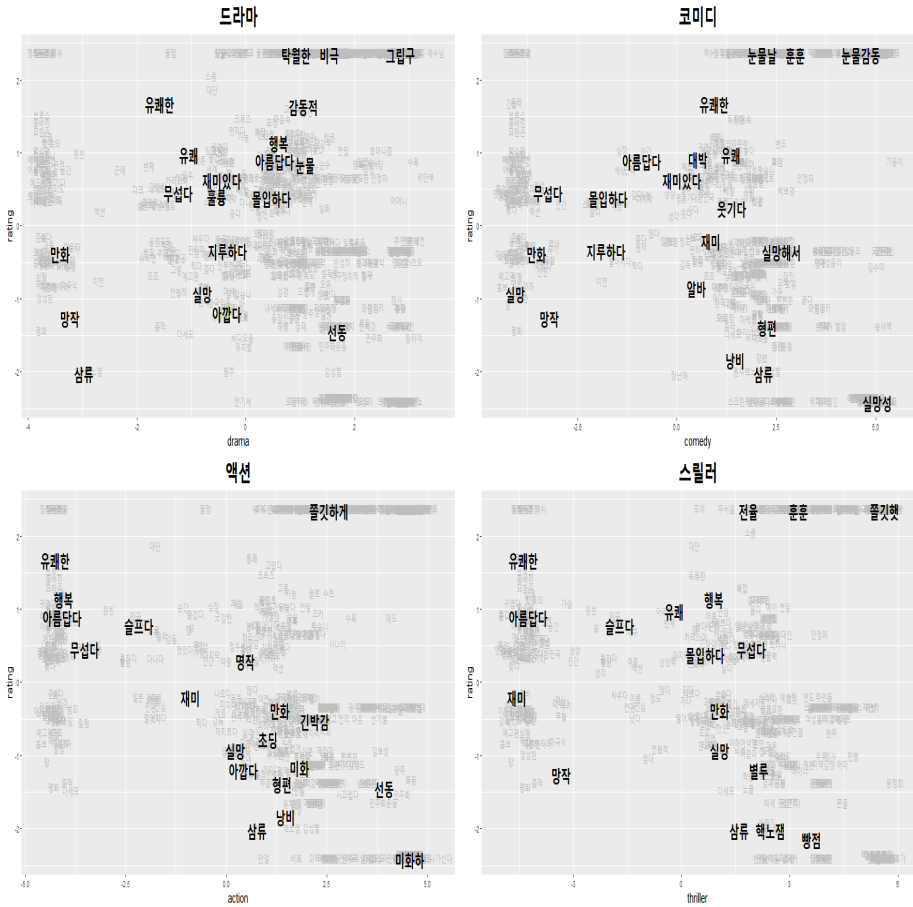

Figure 3: *Estimated term loadings for genre – rating :* $\hat{\phi}_{G\ell,j}$ *vs* $\hat{\phi}_j$ *for word* j.

4. 영화 평점 자료 분석

4.1. 기존 모형 적합 결과

이 절에서는 기존의 다항 역회귀 모형 (3.3)을 적용하여 2장에서 설명한 네이버 영화 평점 자료를 분석한 결과를 설명한다. 식 (3.3)에서, 2.2절에서 설명한 전처리 과정 후 아이디 i의 영화평 텍스트 빈도벡터와 평점을 $\boldsymbol{x}_i$와 $\boldsymbol{v}_i = y_i$로 각각 잡았고, R함수 'mnlm'을 통해 모형 적합 결과를 얻을 수 있다. 8GB RAM의 Intel Core i5-8250U(1.60 GHZ) 환경에서 $18,920 \times 1$ 크기의 모수를 한 번 추정하는데 소요된 시간은 약 80초였다. Figure 1을 통해 총 영화평에서 나오는 단어 빈도수 $\sum_i x_{ij}$와 단어의 역회귀 계수 추정치 $\hat{\phi}_j$를 보여준다. Figure 1은 '재미있다', '감동적', '유쾌한' 등 긍정적인 단어들은 양수인 $\hat{\phi}_j$를 가지며 상대적으로 높은 빈도수를 가지므로 평점이 높을 때 많이 등장하고, '아깝다', '실망' 등 부정적인 단어들은 음수인 $\hat{\phi}_j$를 가지며 상대적으로 높은 빈도수를 가지므로 평점이 낮을 때 많이 등장하는 경향이 있다는 것을 보여준다. '형편','개연성' 과 같은 단어의 의미는 중립적이지만 흔히 '-없다'와 결합되어 부정적인 의미로 사용되어, 음수인 $\hat{\phi}_j$를 가지는 것을 볼 수 있다. 또한, '전율', '개명작', '빵점', '낭비' 등 강한 긍정 혹은 부정을 나타내는 단어들은 절댓값 $|\hat{\phi}_j|$이 매우 크면서 상대적으로 낮은 빈도수를 가지므로 이러한 단어들은 긍적 혹은 부정적인 의견을 강하게 표현하는

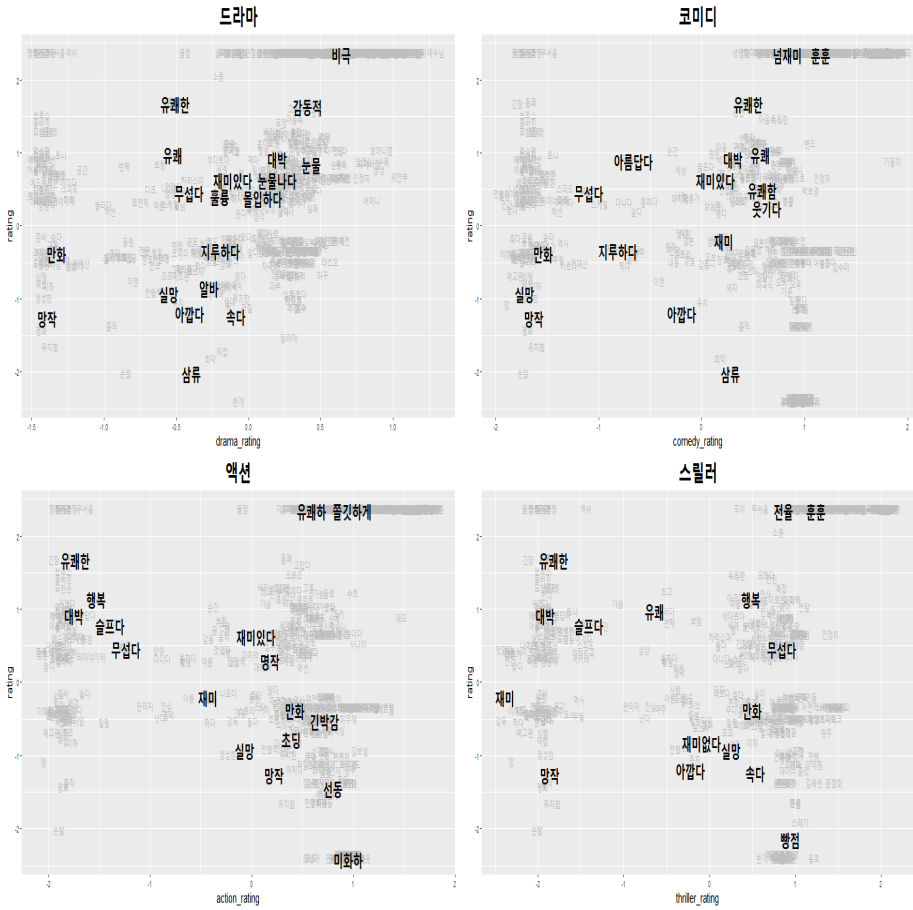

Figure 4: *Estimated term loadings for genre*rating – rating* $\hat{\phi}_{R*G_{\ell},j}$ *vs* $\hat{\phi}_j$ *for word* j.

단어지만 다른 단어들에 비해 많이 사용되지는 않는다는 사실을 보여준다.

4.2. 제안 모형 적용 결과

이 절에서는 3.3절에서 제안한 방법을 적용하여 2장에서 설명한 네이버 영화 평점 자료를 분석한 결과를 설명한다. 4.1절에서처럼 아이디 i의 전처리된 영화평에서 생성된 빈도벡터 $\boldsymbol{x}_i$와 평점 $\boldsymbol{v}_i = y_i$을 이용했고, 추가적으로 영화의 장르속성 G(예를 들어, G_1은 드라마, G_2는 코미디, G_3는 액션, 그리고 G_4는 스릴러)를 사용했다. 다음의 그림들은 제안한 모형으로부터 계산된 단어 로딩 추정치 $\hat{\phi}_j$, $\hat{\phi}_{G_1,j}$, $\hat{\phi}_{G_2,j}$, $\hat{\phi}_{G_3,j}$, $\hat{\phi}_{G_4,j}$ 를 보여준다. Figure 3에서 장르속성에 따라 다르게 설정된 단어로딩 추정치 $\hat{\phi}_{G_\ell,j}$ ($\ell = 1, 2, 3, 4$)를 수평축으로 잡았으며, 수직축은 $\hat{\phi}_j$이다. Figure 2는 단어 j의 라플라스 분포의 비율모수(즉, 페널티 상수)값에 따라 계수 추정치 $\hat{\phi}_{G_1,j}$, $\hat{\phi}_{G_2,j}$, $\hat{\phi}_{G_3,j}$, $\hat{\phi}_{G_4,j}$가 어떻게 변하는지 일부 단어를 예시로 추출하여 나타낸 그래프이다. 여기에서, 녹색, 파랑, 오렌지, 보라색 선은 $\hat{\phi}_{G_1,j}$, $\hat{\phi}_{G_2,j}$, $\hat{\phi}_{G_3,j}$, $\hat{\phi}_{G_4,j}$의 자취해를 의미하며, 수직 점선으로 표시된 페널티 상수(로그)값은 디폴트로 지정된 모형 선택 기준인 AIC에 의해 최종 선택된 모형을 의미한다. 그림들로부터, 장르 속성에 따라서 용어에 대한 추정된 역회귀 계수의 부호와 값이 다름을 확인할 수 있다. 예를 들어, Figure 2로부터 '스릴러'는

Table 3: Selected variables and estimated average of true positive rates (ATPR) for three classes

No. of variables	Variables selected on the best combination	ATPR
1	rating	0.4211
2	rating, thriller_rating	0.4282
3	rating, drama, thriller_rating	0.4283
4	rating, drama, comedy, thriller_rating	0.4236
5	rating, drama, comedy, action, thriller_rating,	0.4256
6	rating, drama, comedy, action, thriller, thriller_rating	0.4224
7	rating, drama, comedy, action, thriller, drama_rating, thriller_rating	0.4242
8	rating, drama, comedy, action, thriller, drama_rating, comedy_rating, thriller_rating	0.4215
9	rating, drama, comedy, action, thriller, drama_rating, comedy_rating, action_rating, thriller_rating	0.4216

스릴러 장르에서 양수 값의 계수 추정치를 가지나 나머지 장르에서는 전부 음의 값을 가지며, Figure 3로부터 단어 '유쾌한'은 코미디 장르에서 양수 값의 계수 추정치를 가지나 다른 장르에서는 모두 음의 값을 가짐을 확인할 수 있다. Figure 4는 영화 장르와 영화평의 교차항에 해당하는 로딩 추정치 $\hat{\phi}_{R*G_\ell,j}$ $(\ell = 1, 2, 3, 4)$ 를 $\hat{\phi}_j$와 비교하여 보여주는데, 영화 장르와 영화평의 교차항에 해당하는 로딩 추정치 $\hat{\phi}_{R*G_\ell,j}$들로부터 장르 별로 강한 긍정 혹은 부정을 나타내는 단어의 사용이 다르다는 것을 확인할 수 있었다. 예를 들어, 드라마 장르에서는 '눈물', '감동적', '슬프다' 등의 단어가 긍정적 의미를 띄는 단어이며, 코미디 장르의 경우 역시 '재미있다', '유쾌한', '환상' 등의 단어가 긍정적인 평점에서 사용되는 것을 알 수 있다.

4.3. 평점 예측 성능 비교

4.3.1. 로지스틱 회귀(logistic regression)

이 절에서는 기존 모형과 제안 모형을 비교하기 위해 각 역회귀로부터 구한 충분축소 점수들을 설명변수로 활용하여 리뷰의 평점을 예측하였다. 리뷰의 평점을 분류하기 위해서, 다중 로지스틱 회귀를 활용한 분류(classification)모형을 택했다. 다시 말해, 기존 다항 역회귀에서 생성된 충분축소 점수 $\hat{\Phi}_R^\top \boldsymbol{f}_i$를 설명변수로 사용해 로지스틱 모형을 적용했고, 또한, 제안 방법에서 생성되는 9개의 충분축소 점수 $\hat{\Phi}_R^\top \boldsymbol{f}_i$, $\Phi_{G_1}^\top \boldsymbol{f}_i$, $\Phi_{G_2}^\top \boldsymbol{f}_i$, $\Phi_{G_3}^\top \boldsymbol{f}_i$, $\Phi_{G_4}^\top \boldsymbol{f}_i$, $\hat{\Phi}_{R*G_1}^\top \boldsymbol{f}_i$, $\hat{\Phi}_{R*G_2}^\top \boldsymbol{f}_i$, $\hat{\Phi}_{R*G_3}^\top \boldsymbol{f}_i$, $\hat{\Phi}_{R*G_4}^\top \boldsymbol{f}_i$를 전체모형(full model)으로 활용한 로지스틱 회귀를 생각했다. 고려한 방법의 예측성능을 평가하기 위하여 자료의 90%를 훈련자료로 나머지 10%를 테스트데이터(test data)로 사용하였다. 모형 선택을 위해 최량부분집합선택(best subset selection)을 고려하였다. 전체 리뷰에서 약 66%의 리뷰가 긍정적인 반면 중립적인 리뷰와 부정적인 리뷰는 약 18%, 16%인 점을 고려해 36개의 하위 모형중 감성 별 민감도(sensitivity) 5-폴드 교차검증 추정치의 평균이 가장 큰 모형을 최종적으로 선택하였다. 모형 선택의 결과는 Table 3를 통해 확인할 수 있고, $\hat{\Phi}_R^\top \boldsymbol{f}_i$, $\Phi_{G_2}^\top \boldsymbol{f}_i$, $\hat{\Phi}_{R*G_4}^\top \boldsymbol{f}_i$ 3개의 충분축소 점수를 설명변수로 사용한 모형이 최종적으로 선택되었다.

추가적으로, 혼동행렬(confusion matrix)을 사용하여 방법들의 예측력을 측정했다. 혼동행렬은 잘못된 예측의 영향력을 알아보기 위해 실제 값과 예측 값의 일치 여부를 행렬로 나타내는 모형 평가 기법이며, 혼동행렬의 열은 예측된 값을 뜻하며 행은 실제 값을 의미한다. 기존 다항 역회귀 방법과 제안한 두 방법의 혼동행렬은 Tables 4–6에 각각 나와 있다. 기존 방법과 제안한 방법의 예측 정확도가 유사하지만, Table 7에서와 같이 각 카테고리(1: 부정, 2: 중립: 3: 긍정)별 민감도와 F1지표를 기준으로 비교했을 때, 제안한 방법(축소모형)은 영화의 장르를 고려하지 않은 MNIR 보다 부정적인 리뷰의 감성을 좀 더 정확히 분류함을 확인할 수 있었다. 기존에 고려한 세 개의 카테고리에서 부정과 중립을 하나의 카테고리로 취급하였을 때 이항분류(binary classification)의 결과를 Table 8에서 확인할 수 있다.

Table 4: Confusion matrix of rating

	1	2	3	**Total**
1	49	97	132	278
2	28	78	188	294
3	41	159	923	1123

Table 5: Reduced model confusion matrix

	1	2	3	**Total**
1	55	90	133	278
2	28	78	188	294
3	43	157	923	1123

Table 6: Full model confusion matrix

	1	2	3	**Total**
1	52	94	132	278
2	31	76	187	294
3	42	161	920	1123

Table 7: Prediction results using logistic regression for each method

분류 카테고리	**MNIR**			**Proposed (reduced model)**			**Proposed (full model)**		
	1	2	3	1	2	3	1	2	3
Sensitivity	0.1762	0.2653	0.8219	0.1978	0.2653	0.8219	0.1870	0.2585	0.8192
F1 Score	0.2474	0.2484	0.7802	0.2722	0.2520	0.7800	0.2581	0.2432	0.779

Table 8: Prediction results using logistic regression for each model (binary classification)

분류 카테고리	**MNIR**		**Proposed (full model)**		**Proposed (reduced model)**	
	0	1	0	1	0	1
Sensitivity	0.4464	0.8178	0.4517	0.8160	0.4499	0.8160
F1 Score	0.4942	0.7797	0.4976	0.7796	0.4961	0.7793

4.3.2. 랜덤 포레스트(random forest)

4.3.1절에서는 장르와 장르와 평점 간의 교차항에 대한 충분축소 점수를 설명변수로 사용해 로지스틱 회귀모형을 이용하여 리뷰의 평점을 예측했다. 그 결과, 리뷰의 평점에 대한 충분축소 점수만을 이용할 때 보다 약간의 분류 정확도의 향상을 확인할 수 있었다. 이 절에서는 장르와 장르와 평점 간의 교차항에 대한 충분축소 점수를 설명변수로 사용하는 것이 비모수적 분류모형의 분류 정확도 향상에 기여할 수 있는지 확인해 보았다. 많이 사용되는 비모수적 분류 모형 중 하나인 랜덤 포레스트를 이용해 리뷰의 감성을 분류하는 것을 고려하였다. 장르와 장르와 평점 간의 교차항에 대한 충분축소 점수의 유의함을 확인하기 위해 평점에 대한 충분축소 점수만을 설명변수로 활용한 배깅 트리(bagged tree)를 비교 방법으로 고려하였다. Table 9은 9개의 충분축소 점수를 설명변수로 사용한 랜덤 포레스트와 평점에 대한 충분축소 점수 $\hat{\Phi}_R^\top \boldsymbol{f}_i$만을 설명변수로 활용한 배깅트리(MNIR)를 비교한다. Table 9을 통해, 장르간의 이질성을 고려하는 것이 분류성능의 상당한 향상을 가져오는 것을 확인할 수 있다.

Table 9: Prediction results using random forest when considering tree-based bagging methods

분류 카테고리	MNIR			**Proposed**		
	1	2	3	1	2	3
Sensitivity	0.2014	0.2313	0.7685	0.2374	0.2891	0.8085
F1 Score	0.2563	0.2115	0.7472	0.3034	0.2588	0.7676

5. 결론 및 논의점

본 논문에서는 한국어로 쓰여진 온라인 영화평 텍스트를 활용하여 영화 평점을 예측하는 분석을 고려하고 있다. 온라인 영화평 텍스트를 단어-문서 행렬로 표현하고 각 영화평에서 평점에 연관있는 단어를 추출하여 예측에 사용한다. 단어-문서 행렬을 이루고 있는 영화평 텍스트의 빈도벡터는 차원이 매우 크기 때문에 예측은 어려운 문제이며, 충분차원축소 기법 중 하나인 다중 역회귀 기법을 활용하여 수만 이상의 차원을 간단하게 요약하여 저차원의 충분축소 점수들로 표현하여 이를 예측에 사용했다. 또한, 영화 장르별 단어의 사용이 다르다는 점에 착안하여 장르에 의존하는 다항역회귀모형을 제안하였다.

본 연구에서는 충분한 정보를 담고 있는 용어들을 토대로 영화 평점을 예측하는 과정을 담았지만 자료에서 리뷰 댓글을 활용하여 리뷰의 추천 및 비추천 수를 예측하는 추가 연구를 진행할 수 있다. 평점 자료 텍스트를 전처리하는 과정에서 용어를 추출할 때 문장 구조를 고려하지 않기 때문에 같은 어순 상의 차이를 파악할 수 없었다. 또한 동음이의어에 대한 사전이 존재하지 않아 단어 의미를 맥락으로 구별하기 힘들다는 한계가 있다. 이는 발전된 전처리 방법으로 해결되어야 할 향후 과제이며, 정밀한 전처리로 분석의 성능 향상을 기대할 수 있다.

이미 기존의 다항 역회귀 기법을 이용한 이항 종속 변수의 분류 예측력은 본 논문과 선행 연구를 통해 알 수 있었다. 추가로 1점부터 10점까지의 평점을 정확하게 예측하는 등 3개 이상의 범주를 가진 종속 변수일 때 예측력 역시 확인할 수 있으며, 이와 같이 의미 있는 단어를 추출하여 다른 분야에 활용할 수 있다. 예컨대, 온라인 상의 기사 댓글이 익명제와 실명제인 경우 의미 있는 용어를 남김으로써 어떤 용어가 댓글에 영향을 주는지 살펴볼 수 있다.

다항 역회귀 기법은 다양한 빅데이터 분석 문제를 설명하는 데 유용할 것이며, 낮은 차원의 정보를 이용하기 때문에 상대적으로 효율적이고 빠른 추정이 가능하다. 이렇게 다중 역회귀와 충분차원축소에 대한 연구가 지속적으로 더 이루어지면 더 좋은 결과를 기대할 수 있다.

References

Adragni KP and Cook RD (2009). Sufficient dimension reduction and prediction in regression, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **367**, 4385–4405.

Cook RD and Weisberg S (1991). Discussion of "sliced inverse regression for dimension reduction", *Journal of the American Statistical Association*, **86**, 28–33.

Feinerer I (2017). Introduction to the tm package text mining in R, Available from: `https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf`

Gupta V and Lehal GS (2009). A survey of text mining techniques and applications, *Journal of Emerging Technologies in Web Intelligence*, **1**, 60–76.

Jeon H (2016). KoNLP: Korean NLP package. R package version 0.80.1, Available from: `https://CRAN.R-project.org/package=KoNLP`

Jeon H (2018). Automatic Korean word spacing with R, Available from: `https://github.com/haven-jeon/KoSpacing`

Jentsch C, Lee ER, and Mammen E (2021). Poisson reduced-rank models with an application to political text data, *Biometrika,* **108**, 455–468.

Li KC (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, **86**, 316–327.

Ma Y and Zhu L (2013). A review on dimension reduction, *International Statistical Review*, **81**, 134–150.

Taddy M (2013). Multinomial inverse regression for text analysis, *Journal of the American Statistical Association,* **108**, 755–770.

Tan KM, Wang Z, Zhang T, Liu H, and Cook RD (2018). A convex formulation for high-dimensional sparse sliced inverse regression, *Biometrika*, **105**, 769–782.

Received December 8, 2022; Revised January 15, 2023; Accepted January 20, 2023

한국어 영화평 댓글 자료 및 다항 역회귀 기법을 활용한 영화 평점 예측 분석

김현진[a], 김지수[a], 이은령[1,a]

[a]성균관대학교 통계학과

요 약

IT 기술의 발달으로 인해 현대 사회에는 온라인 상의 기사, 블로그, 댓글 등 방대한 정보가 유통되고 있다. 이러한 데이터의 대부분은 비정형적인 자료(unstructured data)인 텍스트로 구성되어 있으며 끊임없이 생산되고 있다. 이에 따라 비정형적인 자료를 새롭게 수치화된 자료로 변환하여 텍스트 마이닝을 활용한 분석의 수요가 지속적으로 증가하고 있다. 정형화된 자료로부터 패턴을 찾는 데이터 마이닝(data mining)과 달리 텍스트 마이닝은 이렇게 비정형화된 텍스트로부터 패턴을 찾는다는 점이 가장 큰 차이점이다.

본 연구에서는 고차원의 한국어 텍스트를 최소한의 저차원으로 충분한 정보를 담고 있는지 살펴보는 것을 목적으로 한다. 여러 선행 연구로부터 이미 차원을 축소하는 방법이 소개되었으며, 그 중 다항 역회귀 기법과 충분차원축소 개념을 활용한다. 다항 역회귀 모형 기법을 활용하기 위해서는 리뷰 댓글을 모형에 알맞은 형태로 변환해야 하기 때문에 다양한 방식으로 전처리하고 수치화하였다. 분석 과정에서 요약된 저차원의 정보로 생기는 예측력 등과 같은 문제점을 파악하며 이를 해결하기 위한 방법으로 교차항이 포함된 모형을 제안한다. 또한 차원 축소 방법을 이용하여 기존 모형과 비교 분석하고 최종으로 최소한의 차원으로 예측력과 설명력을 기술한다.

주요용어: 텍스트 마이닝, 충분차원축소, 주변 모형, 다항 역회귀, 다중 반응변수, 감성분석, 영화 평점 예측

이 성과는 한국 연구재단의 지원을 받아 수행된 연구임 (No. NRF-2022R1A2C1012798).

[1]교신저자: (03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: silverryuee@gmail.com