碩士學位 請求論文
指導敎授 李 垠 妗

다항 역회귀와 한국어 리뷰 댓글을 이용한 영화 평점 예측

成均館大學校 一般大學院
統 計 學 科
金 智 秀

碩士學位請求論文

다항 역회귀와 한국어 리뷰 댓글을
이용한 영화 평점 예측

2020

金 智 秀

碩士學位 請求論文
指導敎授 李 垠 姈

다항 역회귀와 한국어 리뷰 댓글을 이용한 영화 평점 예측

Predicting movie ratings using Korean reviews with Multinomial Inverse Regression

成均館大學校 一般大學院
統 計 學 科
金 智 秀

碩士學位 請求論文
指導敎授 李 垠 姈

다항 역회귀와 한국어 리뷰 댓글을 이용한 영화 평점 예측

Predicting movie ratings using Korean reviews with Multinomial Inverse Regression

이 論文을 統計學碩士學位請求論文으로 提出합니다.

2019 年 10 月 日

成均館大學校 一般大學院

統 計 學 科

金 智 秀

이 論文을 金智秀의 統計學

碩士學位 論文으로 認定함.

2019 年 12 月 日

審査委員長

審査委員

審査委員

목 차

표 목차

그림 목차

논문요약

다항 역회귀와 한국어 리뷰 댓글을 이용한 영화 평점 예측

인터넷 기술의 발전으로 인해 방대한 양의 자료가 유통되면서 온라인 상의 댓글 또한 기하급수적으로 증가하고 있다. 이에 따라 텍스트 마이닝 기법에 대해 많이 소개되었지만, 텍스트 마이닝에서 한국어의 경우 영어와 달리 텍스트 처리 기술에 대한 연구가 활발하게 이루어지지 않은 상태이다. 자유로운 어순, 맞춤법 파괴, 용언의 불규칙 사용 등과 같은 특성으로 인해 의미 있는 댓글을 알아내기 위한 어려움이 존재하기 때문이다.

본 연구는 한국어 텍스트로 구성된 고차원의 변수를 축소된 저차원으로 충분한 정보를 담고 있는지 보는 것을 목적으로 한다. 차원을 축소하는 방법으로는 여러 선행 연구가 진행되어 왔으며, 그 중에서 충분 차원 축소와 다항 역회귀 기법을 활용한다. 이러한 모형 기법에 적용하기 적합한 형태로 만들기 위해 리뷰 댓글을 다양한 방식으로 전처리하였다. 이에 저차원의 정보로 생기는 문제점으로 인해 교차항이 포함된 모형을 생성하여 분석하는 방법을 제안하였다. 이에 추가적인 차원 축소 방법을 통해 기존 모형과 비교 분석하여 최종으로 최소한의 차원으로 설명력을 기술한다.

주제어 : 충분 차원축소, 다중 역회귀, 리뷰 평점 예측, 텍스트 마이닝, 감성 분석

제 1장 서 론

1.1 연구 배경 및 목적

현대의 사회에는 IT 기술의 발달로 인해 온라인 상의 기사, 블로그, 댓글 등 방대한 데이터가 발생되고 있다. 대부분의 방대한 양의 정보는 비정형적인 자료(unstructured data)인 텍스트로 저장되어 있으며 이러한 형태의 자료는 끊임없이 생산되고 있다. 이에 따라 비정형 형태인 텍스트를 정형화된 데이터로 새롭게 정의하고 텍스트 마이닝을 통한 빅데이터 분석의 수요가 점점 증가하고 있다. 여기서 텍스트 마이닝은 정보 검색, 데이터 마이닝, 기계 학습, 통계학, 그리고 컴퓨터 언어학 등 여러 분야의 방법을 이용하여 흥미롭고 새로운 정보를 추출하는 방법이다(Gupta와 Lehal, 2009). 정형화된 데이터로부터 패턴을 추출하는 데이터 마이닝(data mining)와는 달리 비정형화된 텍스트로부터 패턴을 추출한다는 점에서 차이가 있다.

텍스트를 머신러닝 방식으로 처리하는 일은 그 자체만으로 충분한 내용을 채울 수 있을 만큼 방대한 주제이다. 더불어 한국어 텍스트를 정밀하게 전처리하여 분석하는 데 의의가 있으므로 전처리 과정에 자세히 서술한다. 본 연구의 최종목표는 분석의 어려움이 많은 한국어 텍스트가 설명변수라고 할 때, 이 설명변수들의 정보를 담고 있는 저차원의 점수로도 감성을 판단할 수 있는지 알아보고자 한다. 텍스트 자료를 충분 차원 축소와 다항 역회귀를 통해 수치형 자료로 변환한 뒤 선형회귀분석을 실시하였을 때 의미가 있는지 역시 살펴본다.

본 연구에서는 분석하고자 하는 자료로 평점이 있는 영화 데이터를 사용하였다. 영화 데이터는 감성 분석의 기본이 되는 자료 중에 하나로, 다항 역회귀 기법을 이

용하여 댓글이 담긴 직관적인 정보를 얻을 수 있다. 또한 기존 다항 역회귀에서 저차원의 정보로 인해 생기는 문제점을 파악하고, 이를 해결하기 위한 방법으로 교차항이 포함된 모형을 제안한다. 이렇게 기존 모형과의 비교 분석을 수행하여 개선점을 기술하고, 분석 결과를 바탕으로 실제 텍스트 자료를 다른 분야에 추가로 응용 및 활용될 수 있는 방향을 제시하고자 한다.

1.2 논문 구성

본 논문은 총 6장으로 구성되어 있다. 제 2장에서는 분석을 실행하기 위한 선행연구들에 대한 내용이 수록되어 있다. 제 3장에서는 실제로 진행한 자료의 전처리에 관한 과정이 들어있다. 제 4장에서는 전처리된 자료를 방법론 모형에 적합한 과정과 결과에 대한 내용이 수록되어 있다. 제 5장에서는 교차항이 포함된 추가 모형을 제안하고 다른 차원 축소 방법들과 비교 결과를 서술한다. 마지막으로 6장에서는 연구 결론 및 논의점에 대해 이야기하며 마무리한다.

제 2장 문헌 연구

본 장에서는 텍스트 자료 전처리 과정의 기본인 TF-IDF(Term Frequency-Inverse Document Frequency) 가중치를 이용한 텍스트 자료의 수치화 변환 및 문서-단어 행렬에 대한 개념을 기술한다. 추가로 수행한 연구 이해를 돕기 위해 차원 축소 방법 중 하나인 충분 차원 축소와 다항 역회귀에 대해 설명한다.

2.1 텍스트 자료

기계학습, 딥러닝을 실행하기 위해서 비정형 데이터인 이미지, 텍스트 등의 데이터는 모두 수치화하여 훈련해야 한다. 데이터 정제 및 전처리는 기계가 텍스트를 이해할 수 있도록 텍스트를 정제하는 과정이다. 그중에서 텍스트를 자연어 처리 모델에 적용할 수 있게 언어적인 특성을 반영해 단어를 수치화하는 작업을 토크나이징(Tokenizing)이라고 한다. 특히 한글의 경우, 띄어쓰기가 잘 되어있으면 단어를 구분하는 토크나이징이 쉬워진다.

2.1.1 문서-단어 행렬과 TF-IDF 가중치

말뭉치(corpus)로부터 단어에 대한 사용 빈도수, 단어 간의 관계, 그리고 문서와 단어 간의 관계를 알기 위해 자연어 텍스트를 정형 데이터의 행렬 방식으로 표현할 수 있다. 이를 문서-단어 행렬(Document-Term Matrix)이라고 하며 각 행은 문서를 나타내고 각 열은 단어를 나타낸다. 원소는 단어가 문서에서 사용되는 빈도를 뜻한다. 텍스트 문서를 수치 데이터로 표현하는 가장 간단한 방법 중에 하나는 용

어빈도(Term Frequency, TF)를 이용하는 방법이다. 문서에서 TF는 단어가 문서에 나타난 횟수를 나타내고, 단어가 문서에 자주 나타나면 그 단어는 문서와 관련이 깊다고 말할 수 있다. 문서-단어 행렬을 전치(transpose)하게 되면 단어-문서 행렬(Term-Document Matrix)이 된다. [표 2.1]은 다음과 같은 문서가 있을 때 문서-단어 행렬로 나타낸 표이다.

	사과(단어 1)	키위(단어 2)	좋아한다(단어 3)	싫어한다(단어 4)
문서 1	1	0	1	0
문서 2	1	1	0	1
문서 3	0	1	1	0

문서 1: 나는 사과를 좋아한다. 문서 2: 나는 사과와 키위를 싫어한다. 문서 3: 나는 키위를 좋아한다.

[표 2.1] 문서-단어 행렬

그러나 이렇게 TF를 이용하여 수치화하는 경우, 문서에서 자주 나타나는 단어 같은 경우는 문서의 의미를 파악하는 데 크게 기여를 하지 않음에도 큰 가중치를 가지게 된다. TF-IDF는 이러한 단점을 해결하기 위해 나온 방법으로, 원래의 TF 값에 단어의 중요도를 고려하여 가중치인 역문서빈도(Inverse Document Frequency, IDF)를 준다. IDF($IDF(d,t)$)는 한 단어가 전체 문서 집합 내에서 공통적으로 얼마나 많이 등장하는지 나타내는 값인 DF($DF(t)$)에서 역수를 취해 준 값이고, 수식으로 나타내면 다음과 같다.

$$IDF(d,t) = \log\frac{N}{1+DF(t)},\ N\text{은 전체 문서의 개수} \quad (1)$$

이렇게 TF-IDF는 식 (2)와 같이 TF 값과 IDF 값의 곱으로 계산된다.

$$(TF-IDF)(t,d) = TF(t,d) \times IDF(t) \quad (2)$$

여기서 t는 단어, d는 문서를 뜻한다. 이 TF-IDF 값이 큰 값이 가지는 단어를 문서에서 높은 식별력을 주기 때문에 정보검색 과정에서 중요한 역할을 하는 단어이다.

2.1.2 정규화된 TF-IDF 가중치

문서의 길이가 길수록 단어의 출현빈도가 높아지게 되고 출현하는 단어의 개수가 많기 때문에 길이가 짧은 문서에 비해 단어가 검색될 가능성이 높다. 따라서 문서의 길이에 대한 영향을 받지 않게 하기 위해 TF에 대한 정규화(normalization)가 필요하다. 정규화된 TF 값은 다음과 같이 나타낼 수 있다.

$$\text{Normalized } TF = \frac{(1+\log_2(TF))}{n_i} \quad (3)$$

정규화된 TF-IDF는 정규화된 TF 가중치에 IDF 값을 곱함으로써 얻어질 수 있다. 이 값은 추후에 분석 모형에서 문서-단어 행렬의 값으로 입력된다.

2.1.3 희소행렬(Sparse Matrix)

문서-단어 행렬은 자연어의 특수성에서 비롯된 특징으로 인한 단점을 가진다. 수만 개가 넘는 단어가 자연어에서 처리될 때, 하나의 문장이나 문서에서 사용되는 단어의 수는 제한적이다. 따라서 이 행렬에서 대부분 칸의 값이 0이 되는데, 이런 종류의 행렬을 희소행렬(Sparse Matrix)이라고 정의한다.

이렇게 0인 값이 많아지게 되면 컴퓨터에서 메모리가 많이 소요된다. 따라서 사용되는 메모리를 줄이기 위해 희소 행렬에 0 요소를 제거하고 오직 비영 요소만을 남겨두는 압축 방식이 쓰인다. 예를 들어 행렬의 95%가 0이라고 할 때, 10만 개의 문서에서 1만 개의 단어를 표현하기 위해서는 대략 20GB가 필요하다. 그러나 이러한 0의 값을 제외한다면 약 150MB 정도로 축약할 수 있다.

2.1.4 워드 클라우드(Word Cloud)

워드 클라우드(Word Cloud)는 텍스트에서 빈번히 사용된 키워드를 시각적으로 표현하는 텍스트 마이닝 방법 중에서 하나로, 단어의 사용빈도 수가 클수록 그 단어를 강조하기 위해 크게 표시된다. 워드 클라우드의 키워드를 시각적으로 표현하게 되면 말이나 글, 혹은 수치로만 설명할 때는 잘 보이지 않는 어떠한 관계의 추세가 드러날 수 있다는 장점이 있다.

2.2 충분 차원 축소(Sufficient Dimension Reduction)

차원 축소(Dimension Reduction)는 어떤 목적에 따라서 데이터 또는 변수의 양을 줄이는 방법이다. 주로 모형의 복잡성(Complexity)이나 과적합(Overfitting)을 피하기 위해 쓰이며 가장 중요한 의의는 관측 데이터를 잘 설명할 수 있는 잠재 공간(latent space)를 찾는 것이다. 다변량 문제를 해결하는 가장 오래되고 잘 알려진 차원 축소의 대표적인 방법은 특이값 분해(Singular Value Decomposition, SVD)와 주성분 분석(Principal Components Analysis, PCA)이다.

통계학에서 충분 차원 축소(Sufficient Dimension Reduction, SDR)는 차원 축소의 아이디어와 충분성(Sufficiency)의 개념을 결합한 데이터를 분석하는 패러다임으

로, 회귀분석에서 독립변수의 차원을 축소하는 데 그 목적이 있다. 어떠한 확률 모형도 가정하지 않는 비모수적 접근인 SDR은 제한적이지 않는다는 점에서 효과적으로 자료를 처리할 수 있다는 큰 장점을 가진다. 일반적으로 SDR은 원회귀(forward regression)보다는 역회귀(inverse regression)을 사용하고, 대표적인 SDR 방법들 중에는 분할 역회귀(Sliced Inverse Regression, SIR)(Li, 1991)와 SAVE(Sliced Average Variance Estimation)(Cook, 1991)이 있다.

반응변수 $y \in \mathbb{R}$이고 예측변수 $\mathrm{X} = (x_1, ..., x_p)^T \in \mathbb{R}^p$인 회귀 문제가 있다고 할 때, 이 회귀의 주된 목적은 $\mathbf{X}$의 값으로 $y|\mathrm{X}$의 조건부 분포 변화를 아는 것이다. 이때 Φ를 고정된 $p \times q (q \leq p)$ 행렬이라고 하고, 다음과 같은 식이 성립된다고 하자.

$$y \perp \mathrm{X} | \Phi^T \mathrm{X} \tag{4}$$

이 방정식은 $\Phi^T \mathbf{X}$이라는 조건하에서 y가 $\mathbf{X}$와 독립이라는 것을 뜻한다. $F(\cdot)$을 분포함수라고 하면, 식 (4)은 다음과 같다고 할 수 있다.

$$F(y|\mathrm{X}) = F(y|\Phi^{\mathrm{T}}\mathrm{X}) \tag{5}$$

(5)는 $y|\mathrm{X}$의 분포가 $y|\Phi^T\mathrm{X}$와 같다는 것을 나타낸다. 식 (4)과 (5)가 성립되면, 축소된 차수의 예측변수 $\Phi^T\mathrm{X}$는 조건부 분포 $y|\mathrm{X}$에 대한 회귀 정보의 손실 없이 원래의 변수 X를 대체할 수 있다. 따라서 이는 p-차원 예측변수 X가 회귀 정보의 손실 없이 $q(\leq p)$차원 예측변수 $\Phi^T\mathrm{X}$로 감소한다고 말할 수 있다.

충분 차원 축소는 역회귀와 깊은 관련이 있으며(Cook, 2005) 역회귀에 대한 자세한 내용은 다음 절에서 다룬다.

2.3 다항 역회귀(Multinomial Inverse Regression, MNIR)

단순 선형 회귀분석(Simple Linear Regression)은 선형적인 상관성을 가진 독립변수와 종속변수 간의 관계를 증명하는 방법이다. 이렇게 텍스트-감성 예측에 대한 단순한 접근방식은 $y_i|x_i$인 일반적인 회귀분석에 적합하는 것이다. 그러나 텍스트 자료가 매우 높은 차원이라는 것을 감안할 때, x_i를 단순화하기 위한 조치를 취하지 않고는 이 조건부 분포를 효율적으로 추정할 수 없다. 따라서 본 연구에서는 단순 선형 회귀분석이 아닌 역회귀분석을 통해 텍스트 데이터를 처리하고자 한다. 다항 역회귀는 x_i가 텍스트로 되어있는 문서라고 하고, 종속변수 y_i를 예측할 때 다항분포에서 비롯된 것으로 나타낼 수 있는 예측변수 세트를 단순화하기 위한 일반적인 방법으로 소개되었다(Taddy, 2013). 더불어 문서 주석에 대한 구문의 로지스틱 회귀는 감성 정보(sentiment information)가 풍부한 저차원 표현을 얻기 위해 사용될 수 있다.

Taddy(2013a)은 주어진 감성에 대한 역조건 분포(Inverse Conditional Distribution)를 통해 y_i와 관련된 정보를 보존하는 저차원 문서 점수를 얻는 역회귀 접근법을 제안했다. 기본적인 다항 역회귀 모형은 다음과 같이 정의된다.

$$\mathrm{x}_i \sim \mathrm{MN}(m_i, q_i) \text{ with } q_{ij} = \frac{\exp(\eta_{ij})}{\sum_{i=1}^{p}\exp(\eta_{ij})}, j = 1,...,p \tag{6}$$

식 (6)에서 각 x_i는 크기 m_i와 확률 q_i를 따르는 $p-$차원 다항분포이며 [1] $\eta_{ij} = \alpha_j + \mathbf{u}_{ij} + \mathbf{v}_i^T\varphi_j$이다.

1) α_j는 절편항이며 $\mathbf{u}_{ij}$는 임의 효과를 뜻한다.

$\mathbf{v}_i$를 $\mathbf{x}_i$가 y_i에 따라 달라지는 반응 인자(response factor)의 K-벡터라고 가정했을 때, Cook(2007)에 의해 고안된 선형 역회귀 공식은

$$\mathrm{x}_i = \Phi \mathrm{v}_i + \epsilon_i \tag{7}$$

이다. 여기서 $\Phi = [\varphi_1, ..., \varphi_K]$는 역회귀 계수의 $p \times K$ $(K << p)$ 행렬이고 ϵ_i는 오차항 p-벡터이다. 일반적으로 v_i는 반응변수가 연속형일 때는 $\mathrm{v}_i = \mathrm{step}(y_i)$로 변환하고, 반응변수가 범주형일 때는 $\mathrm{v}_i = y_i$로 계산한다. $var(\epsilon_i)$의 특정 조건에서 투영(projection) $z_i = \Phi' f_i$은 y_i가 주어진 x_i와 독립이 되도록 충분 축소를 제공한다. 이것은

$$p(\mathrm{x}_i | \Phi' \mathrm{x}_i, y_i) = p(\mathrm{x}_i | \Phi' \mathrm{x}_i) \tag{8}$$

를 의미한다. 이는 데이터인 x_i와 매개변수인 y_i에 대한 충분성의 고전적 정의에 해당하지만, 실제로 추정해야 하는 알려지지 않은 Φ를 조건으로 한다. 이렇게 추정이 가능한 경우, 차수를 p에서 K로 줄인 충분 축소 투영(Sufficient Reduction projection)이 원래 예측변수보다 더 쉽게 작업할 수 있게 된다. 또한, 충분 축소 점수(Sufficient Reduction Score, SR Score)는 다음과 같이 관계 정의된다.

$$y_i \perp x_i | z_i = \Phi' f_i \tag{9}$$

여기서 $f_i = \frac{x_i}{m_i}$는 i번째 문서 안에서 단어의 빈도 비율 벡터를 의미하며,

$m_i = \sum_{j=1}^{p} x_{ij}$는 i번째 문서에서 단어들의 총 빈도수를 뜻한다. 이렇게 충분 축소 투영을 하면, 모든 x_i가 무시되며 텍스트-감성 관계 모형을 적합하는 것이 단일 회귀분석 문제로 바뀌게 된다. 예를 들어, 선형 회귀인 $E(y_i) = \beta_0 + \beta_1 z_i$, 2차 다항식(quadratic)인 $E(y_i) = \beta_0 + \beta_1 z_i + \beta_2 z_i^2$, 그리고 로지스틱 회귀인 $P(y_i < a) = (1 + \exp[\beta_0 + \beta_1 z_i])^{-1}$를 사용할 수 있다. 이 SR 점수는 회귀분석, 분류 등 추가 분석에서도 널리 쓰일 수 있다.

(6)과 같은 수많은 모수가 포함된 식의 모형을 추정할 때, 각 계수 φ_j에 대해 [2] 팻-테일(fat-tailed) 분포와 희소성(sparsity)를 유발하는 독립적인 라플라스 사전 분포(independent Laplace priors)를 사용한다. 변수별 정규화의 적절한 수준에 대한 불확실성을 설명하기 위해 각 라플라스 매개변수를 감마 초사전분포(hyperprior)를 따르는 미지의 값으로 둔다. 예를 들어,

$$\pi(\varphi_j, \lambda_j) = \frac{\lambda_j}{2} e^{-\lambda_j |\varphi_j|} \frac{r^s}{\Gamma(s)} \lambda_j^{s-1} e^{-r\lambda_j}, \quad s, r, \lambda_j > 0, \tag{10}$$

와 같이 Ga(s,r) 초사전분포에 따라 각 j에 대해 독립적이다.

계수와 다른 척도 모수의 최대 사후 확률 추정치(joint posterior maximum, MAP)(Taddy, 2014)를 찾기 위한 새로운 알고리즘이 개발되었다. 알려지지 않은 λ_j 아래의 계수 φ_j를 추정하는 것은 엄청난 비용이 필요로 하는데, 이러한 문제는 패널티(penalty)가 부여된 계수에 대한 로그 최대가능도 문제로 줄여지고, 좌표 하강 방법을 통해 비교적 빨리 해결할 수 있다. 예컨대, (10)의 식과 같이 사전분포가 주어졌을 때 식 (6)에 의해 계산된 로그 가능도는 각 계수에 대한 비용 제약을 적용

2) 극한 상황에서의 확률이 더 높은 분포

하여 최대화된다.

$$c(\varphi_j) = s\log(1+\frac{|\varphi_j|}{r}) \tag{11}$$

이렇게 감마-라쏘(gamma-lasso)라고 불리는 항은 강력하고 새로운 추정이 가능하게 만든다. 이러한 접근법은 수천 개의 고유 토큰을 가진 문서에 대해 아주 짧은 시간 내에 강력한 SR 점수를 산출하기 때문에 계산적으로 매우 효율적이다.

제 3장 분석 자료 및 자료 전처리

3.1 자료에 대한 설명

본 연구에서 사용된 자료는 2016년 1월부터 2016년 12월까지 수집된 140자 평이 있는 [3]네이버 영화 평점 데이터이다.

크롤링된 텍스트 파일을 받으면 기존 변수는 댓글 아이디, 평점, 댓글의 추천 및 비추천, 그리고 리뷰 내용으로 구성되어 있다. 추가로 제목에 쓰여져 있는 장르와 영화 아이디를 불러와 새로운 변수로 생성하였다. 분석에 사용된 변수는 장르, 영화 아이디, 1점부터 10점으로 이루어져 있는 평점, 그리고 리뷰 내용이다. 그 결과 불러온 총 데이터의 관측치 수는 9,541,227개이며, 장르의 종류는 1,583개이다.

장르	영화 ID	평점	리뷰
범죄드라마스릴러	52979	3	범죄자의 성장 드라마
코미디	47531	1	빵점은 없냐
드라마	123298	8	재미남 정말 근데 이거 보러 가기가 힘들었음
스릴러	124025	1	핵노잼 시간 버린 자
범죄드라마	11063	10	희대의 역작을 만났구나

[표 3.1] 영화 평점 자료 예시

3.2 분석 절차

형태소 분석을 하기 위해 R의 한글 자연어 분석 패키지인 KoNLP(Korean

3) 자료 출처는 github.com/drexly/movie140reviewcorpus 이다.

Natural Language Processing, Jeon 2016)를 사용하여 태깅된(tagging) 품사의 단어를 등록된 사전을 통해 추출하였다. 텍스트 마이닝의 가장 대표적인 `tm` 패키지(Feinerer, 2017)를 사용하여 데이터를 말뭉치로 변형하고, 정제된 말뭉치로부터 문서-단어 행렬을 생성하여 역회귀 모형에 적합하기 위한 형태로 나타내었다. 분석을 진행하기 위한 자료 전처리 과정에 대한 내용은 다음 절 내용에서 상세히 기술하였다.

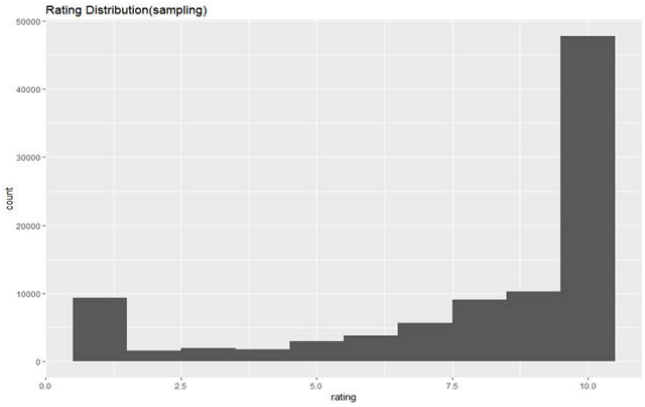
모형에 적합하기 전 분석 결과를 효과적으로 표현하기 위해 사전으로부터 추출된 품사 별 나타난 출현빈도를 바탕으로 워드 클라우드를 구현하여 시각화하였다. 전체 장르 말고도 영화의 특성이 명확하게 드러나는 장르인 드라마, 액션, 코미디, 그리고 스릴러의 워드 클라우드 역시 확인함으로써 대략적인 핵심 키워드를 파악하였다.

이후 문서-단어 행렬을 희소 행렬로 변환한 뒤 다항 역회귀 기법을 적용하여 용어마다 주어진 계수의 값을 확인하였다. 또한 계수 절댓값이 높게 나온 단어들을 추출하여 원래 자료에서의 리뷰 형태와 평점이 어떻게 나왔는지 비교하였다. 마지막으로 다항 역회귀 모형을 통해 얻어진 계수를 토대로 충분 차원 점수를 구한 뒤 평점을 예측하기에 충분한 의미를 담고 있는지 추정함으로써 확인하였다.

3.3 자료 전처리 과정

먼저 데이터를 불러오는 과정에서 장르, 영화 아이디, 평점, 그리고 리뷰 내용이 결측치인 경우 분석 대상에서 제외하였다. 그 결과 약 12만 개의 관측치가 삭제되었다. 이후 모형에 적합하기 위한 훈련 자료를 얻기 위해 1% 샘플링을 하였고, 94,175개의 관측치가 추출되었다. 또한, [그림 3.1]에서 확인할 수 있듯이 평점 중에서도 10점인 점수가 굉장히 많은 불균형한 데이터(imbalanced data)임을 고려해 1

점부터 3점은 1점(부정), 4점부터 7점은 2점(중립), 8점부터 10점은 3점(긍정)인 3개의 감성 카테고리로 다시 변환하였다. 변환 후 부정, 중립, 긍정의 감성을 나타내는 관측치의 개수는 각각 12,902개, 14,133개, 그리고 67,140개이다.



[그림 3.1] 영화 평점 분포

텍스트 마이닝 방법론을 적용하기 위해서는 용도에 맞게 텍스트를 처리하는 작업이 중요하다. 그러나 영어와 달리 어근에 접사가 붙어 의미와 문법적 기능이 부여되는 한국어 같은 경우 텍스트 처리 기술이 많이 발전되어 있지 않은 상태이다. 변화가 크지 않고 공백으로 단어를 구별하는 정도로 토큰화할 수 있는 영어와 다르게 한국어는 명사와 조사를 띄어 쓰지 않고, 용언에 여러 가지 어미가 붙는 등과 같은 이유 때문에 토큰화가 복잡해진다. 따라서 한국어는 더 세심한 작업인 형태소(morpheme) 분석으로 토큰화를 진행하여 단순화할 필요가 있다. 이러한 이유로 형태소 분석부터 한국어 자연어를 처리하는 과정은 상당히 어렵고 많은 시간을 필요로 한다.

한국어의 특성상 띄어쓰기가 잘 되어있지 않아도 의미상으로 전달이 쉽고, 특히 온라인 상에서는 띄어쓰기가 제대로 이루어지지 않고 있다. 그러나 실제로 분석을

함에 있어서는 형태소 분석의 의미가 사라지게 된다. 따라서 띄어쓰기는 형태소 분석 이전에 반드시 수행해야 하는 중요 전처리 과정 중에 하나이다. 그리하여 보다 원활한 분석을 위해 자동 띄어쓰기를 해주는 딥러닝 모델 R 패키지인 `KoSpacing`(Jeon, 2018)을 사용하였다.

KoNLP 패키지 안에서 국립국어원에서 제공하는 세종(Sejong)사전(8만개), 그리고 한국정보화진흥원에서 제공하는 형태소 사전 NIA사전(93만개)을 통해 단어 분류 및 한국어 표현을 분해할 수 있다. 그 외에 사전에 등록되지 않은 단어들을 직접 찾아서 사전에 추가하여 함께 분석에 사용하였다. 기존의 형태소 분석 같은 경우 대부분 명사만 추출하여 분석하는 경우가 많지만, 명사만 사용했을 경우 다른 의미가 있는 품사들이 제외될 수 있다. 따라서 어절을 분석하여 품사 중에서 명사 하나만 추출하지 않고 [4]보통명사(NC), 형용사(PA), 동사(PV)를 붙여주었다. 분석에는 품사를 확정시키고 각 단어들의 태그를 붙여주는 함수인 'SimplePos22'를 사용하였다. 그중에서 동사 같은 경우, '-다'가 빠진 형태로 추출되기 때문에 동사인 단어들은 동사 형태를 유지하기 위해 추가로 다시 '-다'를 붙였다. 단어의 길이가 기본적으로 두 글자 미만인 경우는 분석에 포함하지 않았지만 단어 길이가 한 글자이지만 "잼", "굿", "짱" 과 같이 출현 빈도가 100 이상이고 그 단어로도 의미가 충분히 있다고 판단되는 단어는 직접 따로 추출하여 합쳤다. 형태소 분석을 한 뒤 리뷰 내용이 없는 경우는 분석에 포함시키지 않았다.

이후 분석에 의미가 없는 리뷰 내의 숫자, 구두점, 특수문자를 처리하고 한국어가 아닌 다른 언어인 영어, 중국어, 일본어 등 역시 분석 대상에서 제외하였다. 또한 자연어 처리 분야에서 단순히 자음(ㅋㅋ), 모음(ㅠㅠ)만 있는 경우 역시 완벽한 단어의 형태가 아니고 한글 불용어(stopwords)이기 때문에 제거하였다. 댓글에서 줄 바꿈이 있는 경우 프로그램 상에서 기호가 출력되기 때문에 직접 삭제하였다. 이렇게 제거했을 시 앞뒤와 글자 사이사이에 공백이 생기는 등 2개 이상의 공백이

4) KAIST 품사 태그셋(한나눔에서 기본적으로 사용하는 카이스트 형태소 태그 집합)

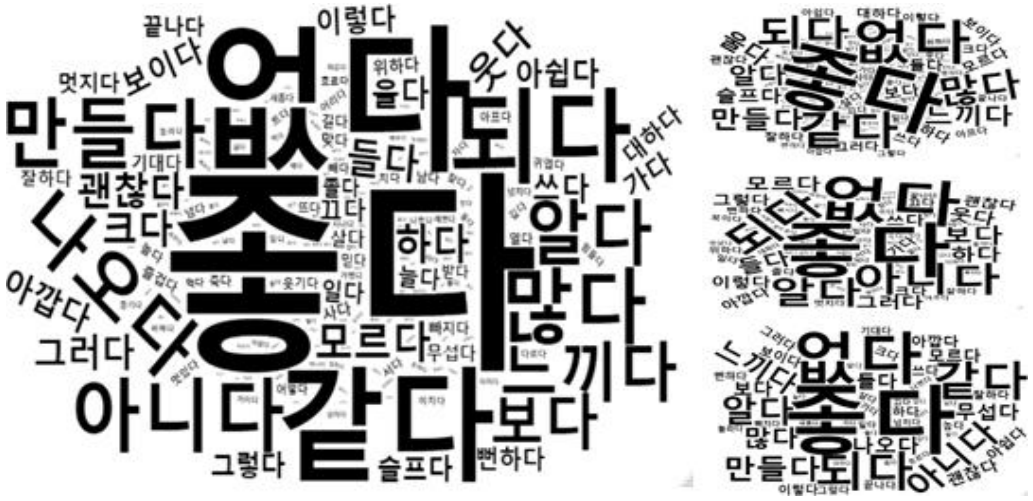
있는 경우가 있어 처리하였다.

이를 바탕으로 워드 클라우드를 통해 전체적인 단어 분포를 파악해본 결과, [그림 3.2]와 같이 나타났다. 보통명사만을 추출하여 살펴보았을 때, 단어 “영화”와 “평점”이 압도적으로 가장 많이 나왔지만, 영화 평점과 직결된 단어이기 때문에 시각화를 표현할 때는 불용어로 취급하였다. 전체 장르에서는 “최고”, “연기”, “진짜”, “스토리” 순으로 출현 빈도가 가장 높게 나왔다. 대표적인 장르 특성을 살펴보기 위해 그림의 우측 상단부터 순서대로 드라마, 액션, 그리고 스릴러 장르 별로 단어 빈도분석 역시 확인해 보았다. 그 결과 드라마 부문에서는 “감동”, “최고”, “눈물”, 액션 부문에서는 “액션”, “재미”, “생각”, “스토리”, 그리고 스릴러 부문에서는 “스토리”, “스릴러”, “반전”과 같은 단어가 가장 많이 출현했다.

[그림 3.2] 워드클라우드(보통명사)

다음 동사와 형용사를 같이 추출하여 살펴보았을 때, 공통적으로 단어 “좋다”의 출현빈도가 가장 높았고 “같다”, “아니다”, “없다”, “되다”와 같은 단순한 보조용언의 비중 역시 높은 것을 확인하였다. 보통명사와 달리 전체 장르와 장르 별로 구분지

어 차이점을 찾기에는 다소 어려움이 있다.



[그림 3.3] 워드클라우드(동사와 형사)

또한 장르의 종류가 1,583개로 상당히 많았던 점을 고려하여 장르 별로 리뷰가 10개 미만인 경우는 대표성이 부족하다고 판단하여 분석 대상에서 제외하였다. 이를 통해 분석에 사용된 장르의 종류는 130개가 되었다. 이후 분석의 효율을 높이기 위해 용어 통일 과정을 수행하였다. 예를 들어, “재미있는”, “재밌는”, “재밌게” 등 의미는 같지만 품사나 어간이 다른 경우와 “재밋는”과 같이 맞춤법이 틀린 경우에는 직접 정규화 표현식을 사용하여 “재미있다”, “재미”와 같이 통일된 단어로 수정하였다.

마지막으로 의미가 없는 리뷰를 가려내기 위해 정규화된 TF-IDF 값을 계산한 뒤 하위 75%에 해당하는 리뷰들을 제거하고, 상의 75%에 해당하는 용어들만 남겼다. 이렇게 모든 전처리 과정을 거친 후 최종적으로 16,973개의 관측치와 18,290개의 단어가 남았다. 문서-단어 행렬로 표현된 차원은 16,973 × 18,290이며 희소성은 100%이다. [표 3.2]는 리뷰의 모든 전처리 과정을 마친 후 분석에 쓰인 최종 자료를

나타낸 것이다.

장르	영화 ID	평점	리뷰(형태소)
범죄드라마스릴러	52979	3(1)	범죄자 성장
코미디	47531	1(1)	빵점
드라마	123298	8(3)	재미있다 힘들다
스릴러	124025	1(1)	핵노잼 시간
범죄드라마	11063	10(1)	희대 역작

[표 3.2] 전처리 후 최종 자료 예시

제 4장 영화 평점 리뷰 분석

4장에서는 다항 역회귀 방법론에 의해 도출된 예측변수를 바탕으로 텍스트-감성 분석에서의 응용을 상세히 기술한다. 모형에 적합하기 위해 사용된 변수와 그에 따라 결과가 어떻게 나오는지 탐색하고, 모형 가정을 충족하기 위한 텍스트 자료 처리 방법 역시 서술한다.

4.1 기존 다항 역회귀 모형 적합 결과

다항 역회귀 방법론에 적용하기 위해 R 패키지 `textir`(Taddy, 2018)를 사용하였다. 그에 앞서 문서-단어 행렬으로 변환된 데이터를 희소행렬로 변환해야 분석 모형에 적절한 형태가 된다. 일반적인 회귀 모형은 이 행렬이 공변량 $\mathbf{x}_i$에 해당하고 y_i가 예측하고자 하는 반응 인자($\mathbf{v}_i$)를 뜻하지만 역회귀 기법에서는 반대로 적용하여 $\mathbf{x}_i$를 추정하기 위한 계수 Φ값을 구한다.

여기서 $\mathbf{v}_i$는 y_i와 동일하게 설정하거나 정수로 반올림하여 계산할 수 있고, 자료 내의 속성을 추가할 수 있다. 속성(attribute)이란 문서 안에서의 추가적인 정보를 뜻한다. 예를 들면, 사용자 아이디나 영화 장르 종류는 자료 내 리뷰가 가지는 특성의 차이를 나타낼 수 있다. 그리고 SR Score을 계산하기 전에 다항 로지스틱 회귀 모형(Multinomial Logistic Regression)의 최대 사후 확률 추정(MAP)을 통해 속성과 계수를 연결할 수 있다.

다항 역회귀를 통해 모형을 적합하면 텍스트와 속성 사이를 연결하는 역할을 하며, 주어진 속성과 직접적으로 관련된 텍스트 내용을 요약하는 충분 축소 점수를

얻을 수 있다.

식 (6)에 의해 각각의 개별 포아송 회귀는 음의 로그우도(Negative log-likelihood)인

$$l(\alpha_j, \varphi_j) = \sum_{i=1}^{n} [m_i e^{\alpha_j + \mathbf{v}_i^T \varphi_j} - x_{ij}(\alpha_j + \mathbf{v}_i^T \varphi_j)] \tag{12}$$

에 비례한다. 더불어 과적합을 방지하고 추정치의 안정화를 위해 각 단어의 계수 크기에 대한 가중 L_1 정규화를 통해 포아송 회귀분석을 다음과 같이 추정한다.

$$\hat{\alpha}_j, \hat{\varphi}_j = \arg\min_{\alpha_j, \varphi_j} \left\{ l(\alpha_j, \varphi_j) + n\lambda \sum_{k=1}^{p} \omega_{jk} |\varphi_{jk}| \right\} \quad \text{where } \lambda, \omega_{jk} \geq 0 \tag{13}$$

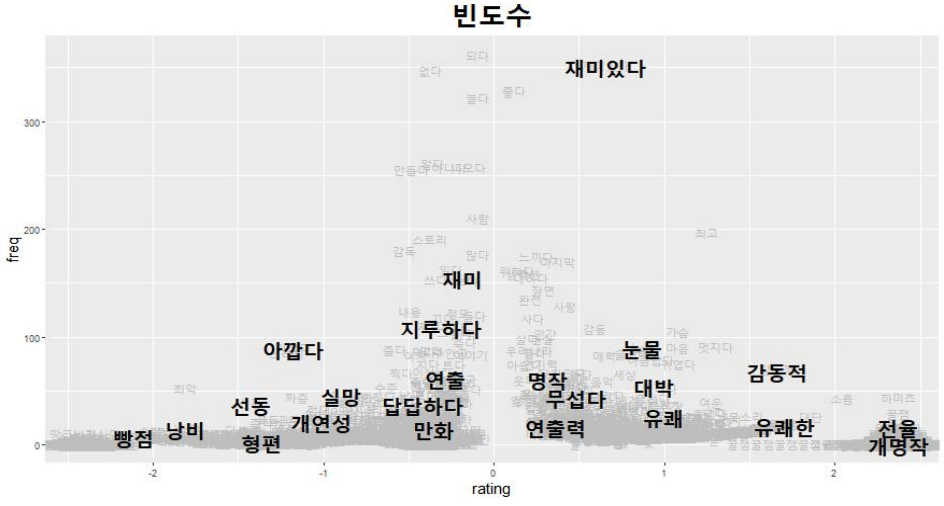
$$\omega_{jk}^t = \left(1 + \gamma \left| \hat{\varphi}_{jk}^{t-1} \right| \right)^{-1}$$

전체 장르를 바탕으로 가장 간단한 모형인 반응 인자를 평점($\mathrm{v}_i = y_i$)을 설정하고, 다항 역회귀 모형에 적합하면 절편항인 α_j값과 계수 φ_j값의 조합인 희소 행렬이 나온다. 더불어, 공변량 항에 4개의 장르 속성(드라마, 코미디, 액션, 스릴러)을 추가한 뒤 모형을 결합하면 장르 속성의 개수만큼 추후에 계산되는 충분 축소 점수 차원의 수가 늘어나게 된다. 초사전분포의 모수 설정은 $s = 1, r = \frac{1}{2}$로 하였고, 성능은 모수 설정 값 변화에 상관없이 매우 로버스트(robust)하다. φ의 값이 결측치인 경우는 희소성을 뜻하며, 희소가 아닌(Non-sparsity) 비율은 [표 4.1]과 같이 나타났다.

	$\hat{\varphi} \neq 0$
평점	0.753
드라마	0.182
코미디	0.078
액션	0.083
스릴러	0.053

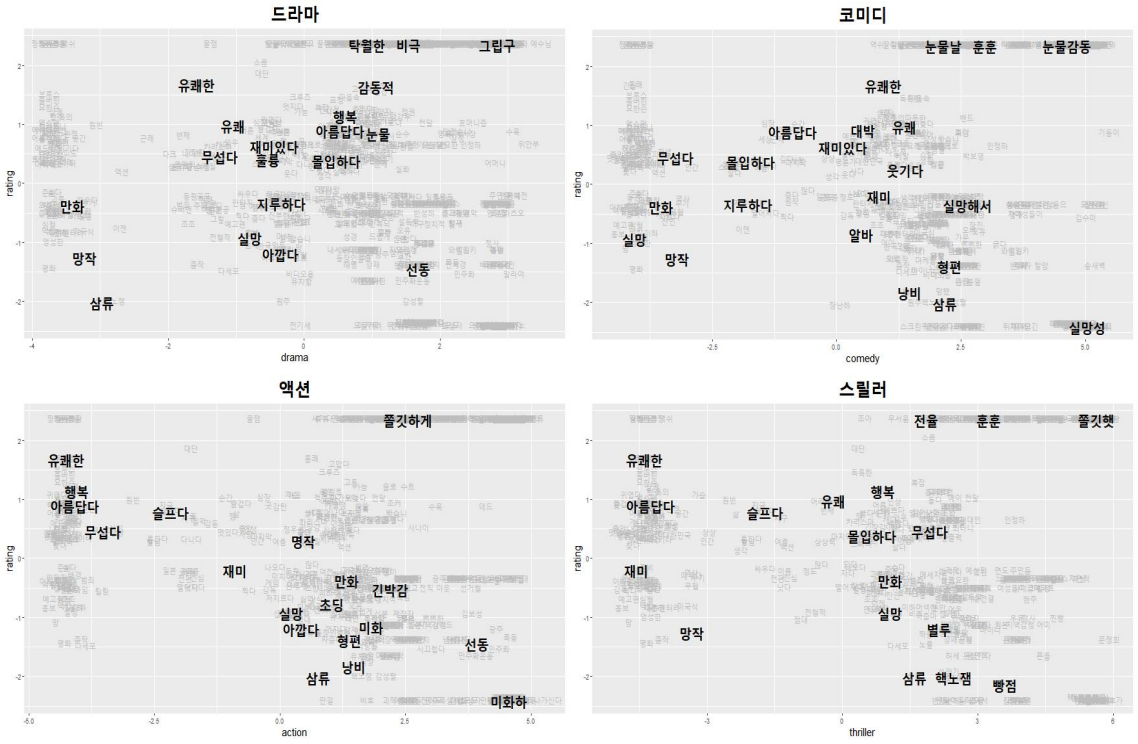
[표 4.1] $\hat{\varphi}$의 Non-sparsity 비율

[그림 4.1]는 문서 안에서 단어의 빈도수와 단어의 고유벡터행렬 Φ(loadings)과의 관계를 그래프로 나타낸 것이다. 드라마와 코미디 장르의 경우 그래프를 살펴보면 단어 “재미있다”와 “재미”가 압도적으로 많이 출현한 반면, φ값이 0에 가까운 것으로 보아 SR 점수에 미미한 영향을 주는 것으로 나타났다. 반대로 높은 음수 값을 가지는 단어 “낭비”, “실망”, “선동”과 높은 양수 값인 단어 “감동적”, “전율”은 자료 내의 출현빈도가 높지 않아도 극값을 가진다. 따라서 빈도수가 많다는 것이 평점 차원 축소 점수가 높게 나온다는 의미를 뜻하지는 않는다.



[그림 4.1] 평점*빈도수의 관계 그래프

[그림 4.2]는 장르와 평점을 나타내는 변수의 관계를 그래프로 나타낸 것이다. 예시로, “행복”이라는 단어는 드라마와 스릴러 장르에서는 양의 극값의 의미를 가지지만, 액션 장르에서는 음의 극값을 가진다. 마찬가지로 단어 “스릴러”는 스릴러 장르에서 양의 값을 가지지만 나머지 장르에서는 전부 음수가 나온다는 사실은 자명하다. 이렇게 똑같은 용어라도 장르 속성에 따라 계수의 값과 부호가 달라지는 것을 알 수 있다.



[그림 4.2] 장르*평점의 관계 그래프

[표 4.2]와 [표 4.3]은 반응 인자 평점의 가장 부정적이고 긍정적인 φ값을 각각 20개씩 뽑았을 때 나오는 용어들을 정리한 것을 나타낸 표이다. 명사 외에도 동사, 형용사 용어가 골고루 나왔음을 확인할 수 있고, 이를 통해 SR Score에 영향을 많이 주는 단어를 알 수 있다.

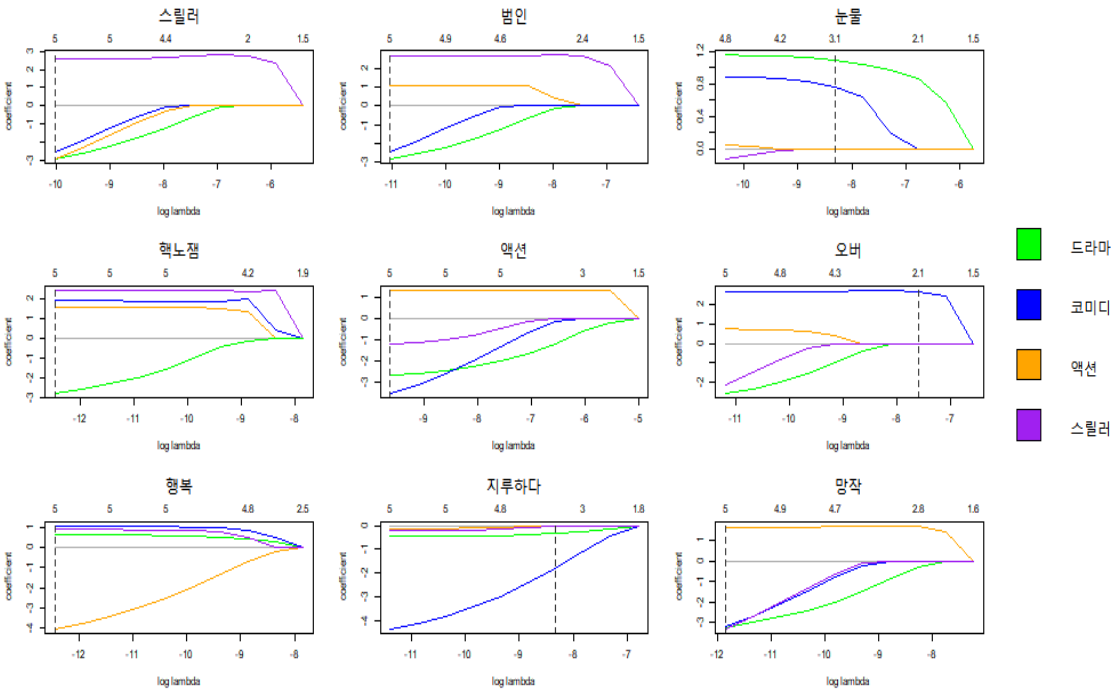
무서움	발랄	전율	황홀	쫄깃	짜릿하다	꿀잼	통쾌한	강추	훈훈
2.44794	2.437734	2.42818	2.4159	2.4031	2.4031	2.3933	2.3819	2.3474	2.3423
생생하다	위대한	소름	긴장	감동적	충분한	상쾌	여운	최고	끌리다
2.3287	2.3142	2.0989	1.6916	1.5749	1.5266	1.2683	1.2580	1.2472	1.2084

[표 4.2] 긍정적인 평점 역회귀계수의 값

손발	삼류	빵점	핵노잼	전라도	희생자	낭비	진압	감성팔	남아돌다
-2.02989	-1.93837	-1.9285	-1.9070	-1.9012	-1.8617	-1.7748	-1.7405	-1.7250	-1.69714
뻐꾸기	망령	메스껍다	난잡함	치욕	최악	망작	노잼	선동	쓰레기
-1.6435	-1.6435	-1.6435	-1.6435	-1.6435	-1.6051	-1.1757	-1.2917	-1.2027	-1.1995

[표 4.3] 부정적인 평점 역회귀계수의 값

`textir` 내의 `mnlm` 함수를 사용하면 감마-라쏘 패널티에 따른 다항 로지스틱 회귀 매개변수를 포아송 가능도에 적합하게 된다. [그림 4.3]은 한 단어의 계수가 반응 인자 별로 어떻게 선택되었는지 예시로 일부 단어를 추출하여 나타낸 그래프이다. 여기서 수직 점선으로 표시된 부분은 지정된 [5]AICc에 의해 선택된 추정치이고, 상단 축은 회귀 자유도를 나타낸다. 이처럼 단어가 속성 변수에 따라 선택되는 값이 달라지는 것을 알 수 있다.



[그림 4.3] 각 단어에 대한 정규화 포아송 회귀계수

5) AICc : $-2l(\hat{\alpha}_j, \hat{\varphi}_j) + 2df_j \frac{n}{n-df_j-1}$ 이며 df_j는 $\{\hat{\alpha}_j, \hat{\varphi}_j\}$에 맞도록 사용된 추정 자유도이다.

4.2 예측을 위한 충분 차원 점수

앞서 구했던 반응 인자를 토대로 지도학습(supervised learning)인 원회귀 분석과 분류 등 다양한 활용 분야에 사용될 수 있다. 다음 표는 차원 축소 투영을 통해 얻은 SR Score 중에서도 제일 높고 낮은 점수의 리뷰 내용 일부분을 나타낸다. 이처럼 텍스트를 토대로 형성된 차원 축소 점수와 평점은 양의 상관관계를 보이는 경향이 있다.

평점	리뷰 내용
10	훈훈하고 이런 영화 보고 싶었는데 딱 좋음
10	애니 영화 중 최고봉인 듯
10	10점 영화 보고 넘 훈훈했어요 짱
10	굿굿 정말 최고
10	아맥으로 봤더니 황홀 그 자체
10	어떠한 호러보다 더 소름 끼치는 영화였다
10	이런 영화 보기 드물다
10	무게감 있는 액션 영화
10	너무 생생한 로봇영화
10	두 번 보고 싶은 영화 정말 인생작
9	심장이 쫄깃
10	간만에 발랄하고 귀여운 영화를 봤다
9	훈훈 오랜 여운
10	통쾌한 풍자
10	굳굳 정말 너무 재밋는 영화네요 추천

[표 4.4] 높은 SR Score을 가진 리뷰 내용

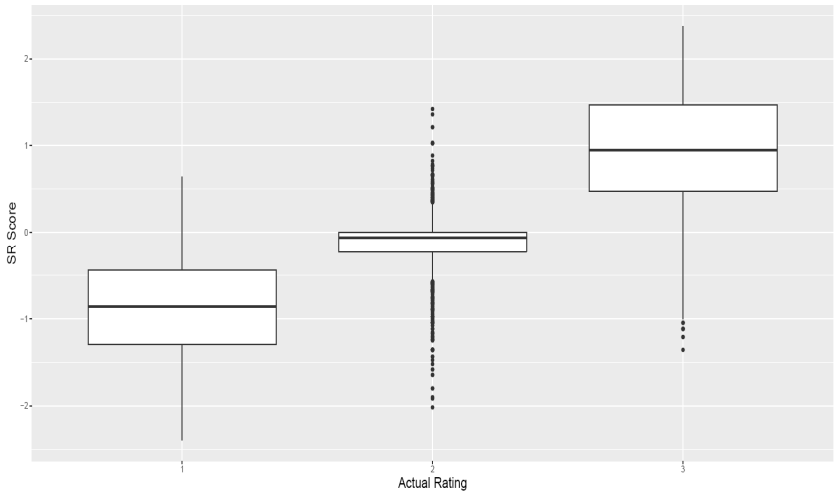
평점	리뷰 내용
1	본 영화 중에 최악중에 최악
3	비호감
1	빵점은 없냐
1	별 반개도 아깝다
1	핵노잼 시간 버린 자
1	손발이 오그라드는 쓰레기 최악중의 최악
1	시간과 돈을 낭비하지 마세요
1	보지마삼 최악임
1	마이너스를 달라
1	환불하고 싶다
3	엄마 죄송합니다 돈 낭비해서
1	폭동을 이런 식으로 바꿔버리네 선동은 정말 무서워
1	아 정말 내가 본 영화 중 손꼽을 만한 최악의 영화였다
1	핵노잼 돈 아까움
1	이거 뭐하는 영화야

[표 4.5] 낮은 SR Score을 가진 리뷰 내용

4.3 영화 평점 예측

다항 역회귀 기법을 통해 얻은 충분 축소 점수를 활용하여 결과변수인 y_i를 다시 추정할 수 있다. 이 절에서는 실제로 추정된 $\hat{\Phi}$가 텍스트의 정보가 축약된 z_i로 충분히 요약이 될 수 있는지 평점을 예측함으로써 확인해보고자 한다.

$\mathbf{v}_i$가 일차원인 경우, 평점을 범주형 변수로 변환하였기 때문에 $v_i = y_i$로 계산하였다. 이를 바탕으로 다항 로지스틱 회귀분석을 이용해 예측 분류를 수행하였다. 다항 로지스틱 회귀분석은 반응변수의 범주가 3개 이상(multivariate)이며 명목형(nominal)일 때 사용하는 대표적인 분류 기법이다. 예측 결과는 정확도(Accuracy)를 통해 성능 비교를 하였다.



[그림 4.4] 실제 평점에 따른 SR Score

상자 그림(box plot)을 보면 긍정적인 리뷰와 부정적인 리뷰를 나타내는 평점은 만 개가 넘는 차원에서 일차원으로 축소된 충분 축소 점수로 충분히 설명할 수 있음을 알 수 있다. 그러나 평점 점수가 4점부터 7점을 뜻하는 중립적인 리뷰는 상대적으로 잘 분류되지 못하고 대부분 긍정적인 평점으로 분류가 되는 경향이 있다.

제 5장 제안 분석 방법 및 실제 적용결과

5장에서는 기존에 적합된 다항 역회귀 모형 외에 교차항(interaction term)을 추가한 다항 역회귀 모형을 제안한다. 이 모형을 사용하여 다중 선형 회귀분석을 실시하고, 최량부분집합 모형과 주성분 분석을 통해 추가로 차원을 축소한 뒤 예측력을 비교한다.

5.1 교차항이 포함된 다항 역회귀

선형 회귀분석에서의 가장 기본적인 목표는 독립변수가 종속변수에 독립적으로 미치는 영향을 회귀계수로 명확히 분석하는 것이다. 그러나 실제 자료에서는 종속변수에 대한 독립적인 주효과(main effect) 외에 상호작용 효과도 분석해야 할 필요성이 있다. 따라서 본 절에서는 다항 역회귀 기법에도 이를 적용시켜 교차항이 포함된 다항 역회귀 모형을 소개한다.

기존의 모형에서 교차항을 더하면 변수 간의 의존성을 알아볼 수 있다. 장르 내의 평점을 고려해야 하기 때문에 평점과 관련된 변수와 교차항을 만들었다. 이렇게 장르 내 속성과 평점 사이에 상호 작용 영향을 알아봄으로써 실제로 평점 예측에 의미가 있는 장르를 알 수 있다. 따라서 장르 속성과 평점의 상호 작용 특성 조합에 대한 교차항을 만들었다. 먼저 앞 절의 분석에서 얻은 충분 차원 점수 $\mathbf{z}_R, \mathbf{z}_{G_1}, \mathbf{z}_{G_2}, \mathbf{z}_{G_3}, \mathbf{z}_{G_4}$는 다음과 같은 기존 역회귀 모형으로부터 얻어졌다.

$$\mathbf{x} \sim \mathrm{MN}(\mathbf{q}(v_R), m) \quad \text{where} \quad \eta_{Rj} = \alpha_j + \mathrm{v}_R \varphi_j \tag{14}$$

$$\mathbf{x} \sim \mathrm{MN}(\mathbf{q}(v_{G_1}, v_{G_2}, v_{G_3}, v_{G_4}), m) \quad \text{where} \quad \eta_{Gj} = \alpha_j + \mathrm{v}_G \varphi_j \tag{15}$$

여기서 R은 평점, G_1는 드라마, G_2는 코미디, G_3는 액션, 그리고 G_4는 스릴러 장르를 뜻한다. 예측하고자 하는 최종 반응변수가 감성 카테고리인 평점인 것을 고려하여 이 변수와 관련된 상호작용을 알아보기 위해 본 분석에서는 총 4개의 새로운 변수를 생성한다. 추가된 추가항을 다음과 같이 정의한다.

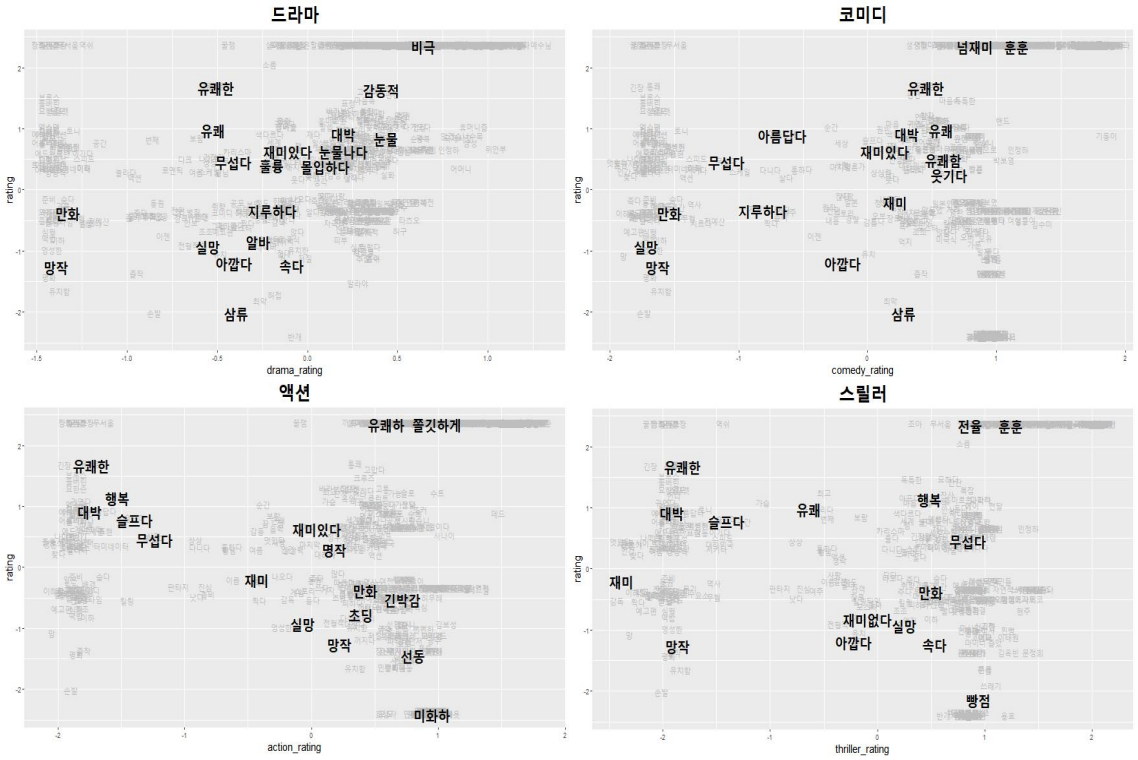
$$\begin{aligned} \mathrm{v}_{R*G_1} &= \mathrm{v}_R \times \mathrm{v}_{G_1} \\ \mathrm{v}_{R*G_2} &= \mathrm{v}_R \times \mathrm{v}_{G_2} \\ \mathrm{v}_{R*G_3} &= \mathrm{v}_R \times \mathrm{v}_{G_3} \\ \mathrm{v}_{R*G_4} &= \mathrm{v}_R \times \mathrm{v}_{G_4} \end{aligned} \tag{16}$$

따라서 기존 다항 역회귀 모형에서 추가된 변수를 포함한 모형은 (16)과 같으며 기존 모형과 결합하여 분석을 실시할 수 있게 된다.

$$\begin{aligned} \mathbf{x} &\sim \mathrm{MN}(\mathbf{q}(v_{R*G_1}), m) \\ \mathbf{x} &\sim \mathrm{MN}(\mathbf{q}(v_{R*G_2}), m) \\ \mathbf{x} &\sim \mathrm{MN}(\mathbf{q}(v_{R*G_3}), m) \\ \mathbf{x} &\sim \mathrm{MN}(\mathbf{q}(v_{R*G_4}), m) \end{aligned} \tag{17}$$

[그림 5.1]은 교차항을 추가한 모형의 변수들 중에서 장르 내의 평점이 평점과 어떤 관계가 있는지 나타낸 그림이다. 기존 모형에서의 장르 별 그래프를 비교해보면 의미 있는 용어가 많이 달라졌음을 확인할 수 있다. 먼저 교차항의 계수의 절댓값은 전체적으로 1을 넘는 경우가 많이 없다. 그러나 장르 내에서 평점에 영향을 확실하게 주는 용어를 알아볼 수 있다. 예를 들어, 드라마 장르 같은 경우, “감동

적", "눈물", "슬프다" 등의 장르의 특성이 드러나며, 코미디 장르의 경우 역시 "유쾌한", "재미있다", "환상" 등의 용어가 나타났다. 새로 생성된 다항 역회귀 계수는 충분 축소 점수로 변환될 수 있으며 추후 절에서 분석으로 쓰인다.



[그림 5.1] 장르 별 평점*평점과의 관계 그래프

5.2 다중 선형 회귀분석(Multiple Linear Regression)

교차항을 포함한 모형을 바탕으로 9개의 설명변수가 예측하고자 하는 반응변수인 평점에 대한 영향력을 살펴보기 위해 다중 선형 회귀분석(Multiple Linear Regression)을 수행하였다.

$$Y=\beta_0+\beta_1\mathbf{z}_R+\beta_2\mathbf{z}_{G_1}+\beta_3\mathbf{z}_{G_2}+\beta_4\mathbf{z}_{G_3}+\beta_5\mathbf{z}_{G_4}+\beta_6\mathbf{z}_{R*G_1}$$
$$+\beta_7\mathbf{z}_{R*G_2}+\beta_8\mathbf{z}_{R*G_3}+\beta_9\mathbf{z}_{R*G_4}$$

이 식에서 Y는 평점을 뜻하며, 각 설명변수는 계산된 충분 차원 점수를 뜻한다. 교차항을 포함한 전체 모형(Full Model)은 [표 5.1]과 같고, 표에서 Estimate은 예측변수에 해당하는 회귀계수를 추정한 값이다.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.162	0.006329	341.611	<2e-16
drama	-0.032453	0.009177	-3.536	0.000407
comedy	0.016743	0.00737	2.272	0.023107
action	0.02222	0.00715	3.108	0.001889
thriller	0.016363	0.006978	2.345	0.0019043
rating	0.570486	0.003897	146.397	<2e-16
drama_rating	-0.008871	0.010842	-0.818	0.41329
comedy_rating	-0.10808	0.012204	-8.856	<2e-16
action_rating	-0.116151	0.010409	-11.159	<2e-16
thriller_rating	-0.06014	0.011936	-5.038	4.74e-07

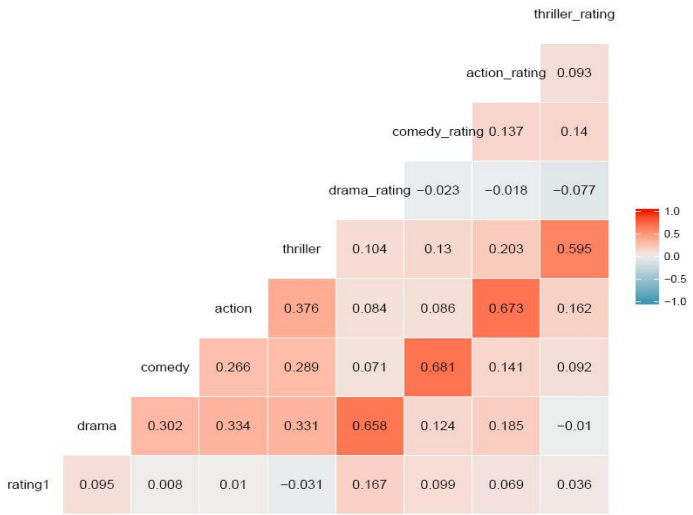
[표 5.1] 교차항이 있는 모형 적합 결과

Pr(>|t|)은 회귀분석 결과의 유의성을 확인하는 지표로, 통계적으로 검정할 때 사용되는 유의확률이다. P-값의 결과를 살펴보면 장르 중에서도 드라마와 평점의 교차항은 유의수준 0.05보다 큰 값을 가지기 때문에 유의하지 않다. 이를 제외한 나머지의 변수는 유의확률이 0에 가까운 값이므로 매우 유의하여 각 변수마다 설명력

을 가지고 있음을 알 수 있다. 적합된 Estimate을 토대로 추정된 회귀모형을 전개하면 다음과 같다.

$$\hat{Y} = 2.162 + 0.57049\,\mathbf{z}_R - 0.03245\,\mathbf{z}_{G_1} + 0.0167\,\mathbf{z}_{G_2} + 0.0222\mathbf{z}_{G_3} + 0.01636\,\mathbf{z}_{G_4} - 0.00887\,\mathbf{z}_{R*G_1}$$
$$- 0.1081\,\mathbf{z}_{R*G_2} - 0.11615\,\mathbf{z}_{R*G_3} - 0.0601\,\mathbf{z}_{R*G_4}$$

중회귀분석을 실시하기 위해서는 설명변수끼리 서로 독립일 것이 요구된다. 이러한 이유로 설명변수 간의 상관관계를 검토하여 설명변수의 편성이 있는지, 즉 상관계수가 ±1에 가까운지 확인해야한다. 다음은 상관분석을 실시하여 추가된 교차항까지 포함한 변수들 간의 상관관계를 행렬로 나타낸 것이다. 가장 높은 상관관계를 보이는 변수는 상관계수가 0.681인 코미디와 코미디 평점의 교차항이다.



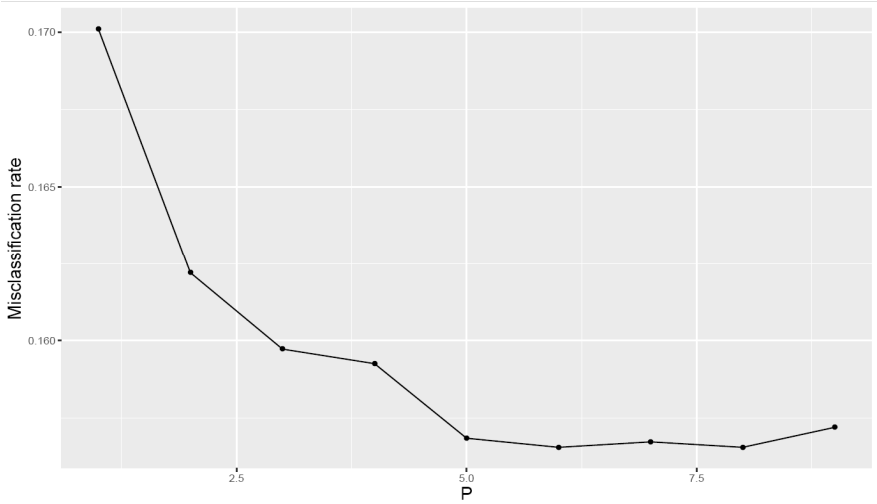
[그림 5.2] 변수 간 상관관계 그래프

5.3 최량부분집합 모형(Best Subset Model)

최량부분집합은 전체 모형에서 보다 많은 모형을 보기 위한 방법으로, 이 기능을 통해 여러 개의 모형을 동시에 봄으로써 선택의 폭을 넓게 가져갈 수 있다. 훈련 자료의 모든 가능한 부분집합 모형의 조합을 살펴보고, 각 변수 개수의 조합마다 오분류율(Misclassification Rate, MCR)이 가장 작은 조합을 선택하도록 하였다. 각 변수 개수의 조합은 [표 5.2]에서 확인할 수 있다. 결과를 요약하면, 1개의 변수(rating)가 선택되었을 때 오분류율이 가장 높았으며 6개 변수의 조합(rating, action, thriller, drama_rating, comedy_rating, action_rating)이 선택되었을 때 가장 낮았다.

변수 개수	부분집합 변수	오분류율
1	rating	0.1700937
2	rating, action_rating	0.1621988
3	rating, drama_rating, action_rating	0.1597243
4	rating, thriller, drama_rating, action_rating	0.1592529
5	rating, action, drama_rating, comedy_rating, action_rating	0.1568373
6	rating, action, thriller, drama_rating, comedy_rating, action_rating	0.1565427
7	rating, comedy, action, thriller, drama_rating, comedy_rating, action_rating	0.1567784
8	rating, drama, comedy, action, drama_rating, comedy_rating, action_rating, thriller_rating	0.1565427
9	rating, drama, comedy, action, thriller, drama_rating, comedy_rating, action_rating, thriller_rating	0.1571908

[표 5.2] 부분집합 변수 개수 별 오분류율



[그림 5.3] 변수 조합별 오분류율 비교 그래프

5.4 주성분 분석(Principal Component Analysis, PCA)

주성분 분석은 고차원 데이터에서 최대 분산의 방향을 찾아 원래의 것과 동일하거나 저차원의 새로운 하위 공간에 투영하는 것을 목표로 한다(Jollife, 2011). 이를 실시하여 보다 작은 차원으로 종속변수를 예측하기 위한 정보를 충분히 담고 있는지 확인할 수 있다. 기존 변수들을 사용하여 주성분을 선택하는 방법 중에 하나는 자료에서 스크리 산점도(Scree plot)를 그려 기울기가 급격하게 변하는 구간을 기점으로 선택하는 것이다. 이를 통해 변수들의 선형결합으로 표시되는 주성분을 추출하여 변수를 축소할 수 있다. [표 5.3]은 누적 변동 비율(설명 비율)을 나타내며, 5개의 중요한 주성분으로 전체 변동(variance)의 약 87% 정도를 설명한다. 즉, 9개의 원래 예측변수에서 주성분 5개로 줄여 추가 분석을 진행할 수 있다.

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.6257	1.2418	1.1784	1.1031	0.9985
Proportion of Variance	0.2937	0.1741	0.1543	0.1352	0.1108
Cumulative Proportion	0.2937	0.4678	0.6221	0.7573	0.8680

[표 5.3] 누적 변동 비율

5.5 기존 모형과 제안 모형의 성능 비교

앞의 결과로 기존 모형과 교차항이 포함된 예측 분류를 다중 로지스틱 회귀를 통해 비교하였다. 예측력을 비교하기 위해 모형의 성능을 평가할 때 사용되는 지표 중에 하나인 혼동행렬(confusion matrix)을 활용하였다. 혼동행렬의 행은 실제 값을 뜻하며, 열은 예측된 값이다. 반응변수가 평점 하나만 있는 경우와 장르 속성이 포함된 모형은 정확도가 83% 정도로 두 모형은 굉장히 유사한 예측력을 나타내는 것을 알 수 있다. 또한, 제안 모형인 전체 모형(Full Model)과 축소 모형(Reduced Model)의 혼동행렬을 살펴보면, 단순 정확도에서는 차이가 없어 보이지만, 중립적인 평점의 카테고리의 정확도가 다소 향상이 있는 것을 확인할 수 있다.

각 카테고리 별로 정확도와 정밀도(Precision)와 재현율(Recall)의 평균인 F1 점수를 비교한 결과를 [표 5.8]로 나타내고, 다중 분류와 이진 분류 결과의 차이를 알아내기 위해 추가적으로 [표 5.9]를 통해 성능 평가를 실시하였다.

	1	2	3	합계
1	2067	566	134	2767
2	303	1378	1266	2947
3	93	525	10641	11259

[표 5.4] 기존 모형 혼동행렬

	1	2	3	합계
1	2073	570	124	2767
2	291	1467	1189	2947
3	87	601	10571	11259

[표 5.5] 장르의 속성이 포함된 기존 모형 혼동행렬

	1	2	3	합계
1	2088	572	107	2767
2	286	1790	871	2947
3	97	735	10407	11259

[표 5.6] 전체 모형 혼동행렬

	1	2	3	합계
1	2089	571	107	2767
2	278	1808	861	2947
3	90	750	10419	11259

[표 5.7] 축소 모형 혼동행렬

	MNIR			MNIR with attributes			MNIR Full Model			MNIR Reduced Model			PCA Model		
분류 카테고리	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
정확도	0.7470	0.4676	0.9451	0.7492	0.4978	0.9389	0.7546	0.6074	0.9243	0.7550	0.6135	0.9254	0.7546	0.4934	0.9265
F1 score	0.7904	0.5089	0.9134	0.7946	0.5253	0.9135	0.7973	0.5923	0.9201	0.80	0.5951	0.9202	0.7904	0.5125	0.9075

[표 5.8] 각 모형의 카테고리 별 분류 비교 결과

각각의 카테고리 별 정확도를 살펴보았을 때 중립적인 평점의 카테고리에서 정확도가 최종 모형에서 15%의 향상이 있는 것을 확인할 수 있다. 불균형한 자료임을 고려하여 F1 점수 역시 살펴보았을 때 약 10% 정도의 높은 성능을 보인다. 반면 긍정적인 리뷰를 1, 나머지를 0으로 이진 분류를 실시하면 기존 모형과 최종 모형에서의 성능은 유사하게 나타났다.

	MNIR		MNIR with attributes		MNIR Full Model		MNIR Reduced Model		PCA Model	
분류 카테고리	0	1	0	1	0	1	0	1	0	1
정확도	0.8612	0.9617	0.8614	0.9618	0.8614	0.9617	0.8630	0.9616	0.8624	0.9598
F1 score	0.8894	0.9465	0.8896	0.9466	0.8895	0.9465	0.8902	0.9468	0.8883	0.9458

[표 5.9] 각 모형의 카테고리 별 분류 비교 결과(이진 분류)

제 6장 결론 및 논의점

본 논문에서는 고차원의 텍스트에서 의미 있는 용어를 파악하는 연구를 진행하였다. 충분 차원 축소와 다중 역회귀 기법을 활용하여 매우 큰 차원을 간단하게 표현할 수 있는 저차원으로 만들었고, 추가적으로 주성분 분석 등을 이용해 차원을 축소하여 최종적으로 5개 이내의 변수로도 충분한 설명력을 가지고 있다는 것을 확인할 수 있었다. 본 연구에서는 충분한 정보를 담고 있는 용어들을 이용하여 영화 평점을 예측하는 과정을 담았지만 추가 연구 과제로 자료에서 리뷰 댓글을 활용한 추천 및 비추천의 수를 예측해볼 수 있다.

영화 평점 자료를 전처리하는 과정에서 단어를 추출할 때 문장 구조를 무시하기 때문에 같은 어순 상의 차이를 고려하지 않았다. 또한, 동음이의어에 대한 사전을 따로 구축하지 않았기 때문에 맥락으로 단어 의미를 구별하기 어렵다는 한계가 있다. 이는 발전된 전처리 방법으로 해결되어야 할 방향이며, 향상된 전처리로 분석의 성능 향상을 기대할 수 있다.

이미 다항 역회귀를 이용한 이항 종속변수의 분류 예측력은 선행 연구를 통해 알 수 있었다. 추가로 1점부터 10점까지 평점을 정확하게 예측하는 등 다중 종속변수일 때의 예측력 역시 확인해볼 수 있다. 이렇게 의미 있는 단어를 추출하여 다른 분야에 활용할 수 있다. 예를 들어, 온라인 기사 댓글이 익명제와 실명제일 때 의미 있는 단어를 추출함으로써 어떤 단어가 댓글에 영향을 주는지 알아볼 수 있다.

다중 역회귀 기법은 다양한 빅데이터 분석 문제를 설명하는 데 유용할 것이며, 상대적으로 빠른 추정과 실험을 할 수 있을 정도로 효율적이다. 이렇게 충분 차원 축소와 다중 역회귀에 대한 연구가 지속적으로 더 발전되면 더 좋은 결과를 기대할 수 있다.

참 고 문 헌

[1] Cook, R.D.(2007). Fisher lecture: Dimension reduction in regression. *Statistical Science 22, 1-26.*

[2] Cook, R.D.(2005). Sufficient Dimension Reduction via Inverse Regression: A Minimum Discrepancy Approach. *Journal of the American Statistical Association.*

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

[4] Gupta, V. and Lehal G. S. (2009). A survey of text mining techniques and applications, *Journal of emerging technologies in web intelligence 1, 60-76.*

[5] Gentzkow, Matthew and Kelly, Bryan T. and Taddy, M. (2017). Text as Data. *Journal of the Economic Literature*

[6] Jeon, H. (2016). KoNLP: Korean NLP package. R package version 0.80.1. https://CRAN.R-project.org/package=KoNLP

[7] Jeon, H. (2018). Automatic Korean word spacing with R. https://github.com/haven-jeon/KoSpacing

[8] Jolliffe, I. T. (2011). Principal Component Analysis. *Springer-Verlag, New York.*

[9] KAIST Semantic Web Research Center (2011). 한나눔 한국어 형태소 분석기 사용자 매뉴얼.
https://www.sketchengine.eu/wp-content/uploads/Original-HanNanum-manual.pdf

[10] Taddy, M. (2013). Multinomial Inverse Regression for Text Analysis.
Journal of the American Statistical Association 108(503), 755-770

[11] Taddy, M. (2013b). Rejoinder: Efficiency and structure in MNIR.
Journal of the American Statistical Association 108, 772-774.

[12] Taddy, M. (2013c). Measuring political sentiment on Twitter: Factor optimal design for multinomial inverse regression.
Technometrics 55, 415-425.

[13] Taddy, M.(2014). One-step estimator paths for concave regularization.

[14] Taddy, M. (2015). Distributed Multinomial Regression. *The Annals of Applied Statistics, 1394-1414.*

[15] Taddy, M.(2018). Package 'textir' for Inverse Regression for Text Analysis.
https://CRAN.R-project.org/web/packages/textir/textir.pdf

ABSTRACT

Predicting movie ratings using Korean reviews with Multinomial Inverse Regression

Kim, Ji Su
Department of Statistics
Sungkyunkwan University

As Internet technology advances have led to the circulation of vast amounts of data, online comments are also growing exponentially. Although there were a lot of studies introduced on text mining techniques, the Korean language has not been actively studied in text mining. Our goal of this study is to see whether the high-dimensional variables contain sufficient information through dimension reduction. To do so, we use the non-parametric approach of Sufficient Dimension Reduction(SDR) and Multinomial Inverse Regression(MNIR). Review comments are pre-processed in a variety of ways to make them suitable for these model techniques. In addition, we will compare the performance of the existing model and the model with interaction term using the other dimension reduction method.

Keywords : Text Mining, Sufficient Dimension Reduction, Multinomial Inverse Regression, Sentiment analysis, Review rating prediction