Journal of the Korean Data & Information Science Society
2016, **27**(4), 845–854

http://dx.doi.org/10.7465/jkdi.2016.27.4.845
한국데이터정보과학회지

온라인 리뷰에서 평점의 분류[†]

최동준[1] · 최호식[2] · 박창이[3]

[13]서울시립대학교 통계학과 · [2]경기대학교 응용정보통계학과

접수 2016년 6월 29일, 수정 2016년 7월 18일, 게재확정 2016년 7월 22일

요 약

감성분석 (sentiment analysis) 혹은 오피니언 마이닝 (opinion mining)은 블로그, 리뷰, 신문기사나 소셜네트워크 등의 문서에서 개인의 주관적인 정보 혹은 의견을 알아보는데 사용되는 텍스트 마이닝의 기법이다. 평점이 있는 온라인 리뷰에서 리뷰 텍스트에 기반한 평점의 분류문제에 대한 선행연구에서는 이진 분류만을 고려하였다. 그러나 긍정과 부정 외에도 중립적인 의견도 있을 수 있기 때문에 이진 분류보다는 다범주 분류가 더 적합할 것이다. 본 연구에서는 리뷰 텍스트에 기반한 평점의 다범주 분류문제를 고려한다. 전처리에서는 카이제곱 통계량을 이용하여 평점과 연관된 단어들을 추출하고 이를 입력변수로 삼아 지지벡터기계 (support vector machines)와 비례오즈 모형 (proportional odds model) 등 다범주 분류기의 예측력을 비교한다.

주요용어: 감성분석, 다범주 분류, 오피니언 마이닝, 워드클라우드.

1. 서론

최근 인터넷의 발달에 따라 다양한 형태의 텍스트 데이터들을 접할 수 있다. 특히 영화나 상품 등에 대한 정보를 보여주는 웹 사이트에서는 이용자들이 영화나 상품들에 대한 리뷰를 작성하고 평점을 매길 수 있는 시스템을 제공하고 있다. 감성분석 (sentiment analysis) 혹은 오피니언 마이닝 (opinion mining)은 블로그, 리뷰, 신문기사나 소셜네트워크 등의 문서에서 개인의 주관적인 의견을 알아보는데 사용되는 텍스트 마이닝의 주요 기법중 하나이다. 텍스트 마이닝을 이용한 여러 가지 흥미로운 분석 사례들은 Bae 등 (2013), Chae 등 (2013), Kim 등 (2013), Lee와 Suh (2014) 등을 참조할 수 있다. 또한 감성분석 전반에 대한 자세한 소개는 Liu (2012)를 참고하기 바란다.

본 연구에서는 평점이 있는 온라인 리뷰에서 리뷰 텍스트에 기반한 평점의 다범주 분류문제를 고려한다. 관련된 선행 연구로는 다음과 같은 것들이 있다. Kim과 Kim (2014)에서는 감성 사전을 구축하여 리뷰 데이터를 분류하였고, Lee와 Hong (2015)에서는 불용어 (stopwords)를 제외한 형용사, 부사, 동사, 명사 등의 4품사 (parts of speech)에 한해 카이제곱 통계량을 이용하여 중요한 단어를 추출한 후 입력변수로 사용하여 분류문제를 고려하였다. 온라인 리뷰의 평점은 보통 순서형으로 다범주인 경우가 많은데 언급한 선행연구들에서는 긍정/부정이라는 두 범주로 이루어진 이진 분류만을 고려하였다.

[†] 이 논문은 2015년도 정부 (교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구 사업임 (No. 2015R1D1A1A01059984).
[1] (02504) 서울특별시 동대문구 서울시립대로 163 (전농동), 서울시립대학교 통계학과, 석사과정.
[2] (16227) 경기도 수원시 영통구 광교산로 154-42 (이의동), 경기대학교 응용정보통계학과, 조교수.
[3] 교신저자: (02504) 서울특별시 동대문구 서울시립대로 163 (전농동), 서울시립대학교 통계학과, 부교수. E-mail: park463@uos.ac.kr

본 연구에서는 온라인 리뷰 텍스트에서 평점을 잘 예측하는 단어들을 추출하기 위해 Lee와 Hong (2015)처럼 카이제곱 통계량을 활용하여 평점과 연관성이 높은 단어들을 추출한다. 또한 추출된 단어들을 입력변수로 이용하여 지지벡터기계 (support vector machine)와 비례오즈 모형 (proportional odds model) 등 여러 가지 다범주 분류 알고리즘을 이용하여 평점에 대한 예측력을 비교한다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 온라인 리뷰에서 카이제곱 통계량을 이용하여 평점과 연관성이 높은 단어들을 추출하는 입력변수의 선택법과 지지벡터기계와 비례오즈 모형을 이용한 다범주 분류법을 소개한다. 3절에서는 모의실험과 실제 온라인 리뷰 데이터에 대하여 여러 가지 지지벡터기계와 비례오즈 모형의 예측력을 비교한다. 마지막으로 4절에서는 본 연구의 결과를 요약하고 후속 연구방향에 대하여 논의한다.

2. 단어 추출법과 다범주 분류법

2.1. 단어 추출법

온라인 리뷰를 비롯한 비정형의 텍스트 데이터에 대한 통계적인 분석에서는 기본적으로 단어의 빈도와 같은 계량적인 변수를 산출하여 정형데이터로 만드는 전처리 과정을 필요로 한다. 단어의 갯수는 문서의 갯수에 비해 기하급수적으로 늘어나게 되므로 적절한 갯수의 단어들을 분석에 이용해야 한다. 평점과 내용을 같이 작성하는 리뷰 데이터에서는 텍스트 뿐 만 아니라 선호도를 나타내는 정량적인 평점이 같이 관측된다. 따라서 단어의 출현빈도와 평점과의 연관성을 반영하여 분석에 사용할 적절한 갯수의 단어들을 추출하여 분석하는 것이 효율적이다.

본 연구에서의 Lee와 Hong (2015)에서 설명한 방식대로 전처리를 하였다. 첫째, 텍스트 데이터들에 있는 불용어들을 제거하고 남는 단어들에 대하여 형용사, 부사, 동사, 명사의 4품사에 대한 정보를 갖는 범주형 변수를 생성하였다. 둘째, 평점 $(= 1, \ldots, J)$과 연관성이 높은 단어들을 추출하기 위하여 Table 2.1과 같이 각 단어를 포함하는 리뷰와 포함하지 않는 리뷰에 대하여 $2 \times J$ 분할표를 작성하고, 카이제곱 통계량을 활용하여 단어의 출현 유무에 따른 분포의 동일성 검정을 수행하여 특정 단어와 리뷰의 평점간의 연관성을 계량화하였다.

Table 2.1 Cross-table for the occurrence of a word in reviews and their ratings

	Rating 1	Rating 2	...	Rating J	Sum
# of reviews with the word	n_{11}	n_{12}	...	n_{1J}	n_{1+}
# of reviews without the word	n_{21}	n_{22}	...	n_{2J}	n_{2+}
Sum	n_{+1}	n_{+2}	...	n_{+J}	n

Table 2.1의 분할표로부터 카이제곱 통계량은 다음과 같이 구할 수 있다 (Agresti, 2002).

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}, \qquad df = J - 1.$$

여기서 $\hat{\pi}_{i+}{=}n_{i+}/n$와 $\hat{\pi}_{+j}{=}n_{+j}/n$는 각 행과 열의 표본 주변비율이며, $\hat{\mu}_{ij}{=}n\hat{\pi}_{i+}$와 $\hat{\pi}_{+j}{=}n_{i+}n_{+j}/n$는 기대빈도이다. 어떤 단어에 대한 카이제곱 통계량이 크면 리뷰에서 그 단어의 출현 여부에 따라 평점의 분포가 동일하지 않다고 볼 수 있다. 이와 같이 카이제곱 통계량을 이용하여 평점과 연관성이 높은 단어들을 추출하여 이후의 분류에서 입력변수로 활용할 수 있다.

2.2. 지지벡터기계

지지벡터기계는 출력값 혹은 클래스값이 +1과 -1로 이루어진 이진 분류에서 마진 (margin)을 최대화하는 초평면 (hyperplane)을 찾는 학습기법이다. 지지벡터에 대한 자세한 소개는 Vapnik (1995)을 참고할 수 있다. 지지벡터기계를 클래스의 갯수가 세개 이상인 다범주 분류로 확장하는 방법은 여러 가지가 있는데, 본 연구에서는 이진 분류의 지지벡터기계를 반복적으로 적용하는 방법만을 고려하기로 한다. 우선 몇 가지 기호를 도입하면 다음과 같다. $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$은 주어진 훈련데이터이다. 여기서 $\boldsymbol{x}_i \in \mathbb{R}^p$와 $y_i \in \{1, \ldots, J\}$는 각각 i번째 입력변수 벡터와 출력변수를 나타낸다.

본 논문에서는 다범주 지지벡터기계 알고리즘으로 OVR (one versus rest), OVO (one versus one), OVN (one versus the next), OVNF (OVN + forward)을 고려한다. 이들중 OVN과 OVNF는 평점의 순서성을 고려하는 방법이고 OVR과 OVO는 평점의 순서성을 고려하지 않는 방법으로 볼 수 있다.

OVR

$j\ (= 1, \ldots, J)$번째 분류기의 새로운 출력변수를 다음과 같이 정의한다. $i = 1, \ldots, n$에 대하여

$$z_i^j = \begin{cases} 1, & y_i = j \\ -1, & y_i \neq j. \end{cases}$$

훈련데이터 $\{(\boldsymbol{x}_i, z_i^j)\}_{i=1}^n$에 대하여 지지벡터기계를 적합하여 얻은 분류함수를 f_j로 나타내자. 새로운 데이터 $\boldsymbol{x}$에 대한 클래스 예측값은 다음의 식을 통해서 얻을 수 있다.

$$\hat{y} = \underset{j=1,\ldots,J}{\arg\max} f_j(\boldsymbol{x}) \tag{2.1}$$

즉, $\hat{y}$는 J개의 분류함수들 중 가장 큰 값을 주는 분류함수에 대응되는 클래스이다.

OVO

OVO 방법은 모든 가능한 클래스값들의 쌍에 대하여 $\binom{J}{2}$개의 쌍별비교를 고려한다. 각 j와 $k(\neq j)$에 대하여 클래스값이 j인 경우 $+1$로 코딩하고 k인 경우 -1로 코딩한 후, 지지벡터기계를 적합하여 얻은 분류함수를 f_{jk}라 하자. 새로운 데이터 $\boldsymbol{x}$에 대한 클래스 예측값은 $\binom{J}{2}$개의 분류함수 $f_{jk}(\boldsymbol{x})$들에 의해 가장 많이 예측된 출력값으로 주어진다. OVO에 대한 보다 자세한 사항은 Hsu와 Lin (2002)를 참고할 수 있다.

OVN

$j\ (= 1, \cdots, J-1)$에 대하여 새로운 출력변수의 값을 다음과 같이 정의하자. $i = 1, \ldots, n$에 대하여

$$z_i^j = \begin{cases} 1, & y_i > j \\ -1, & y_i \leq j. \end{cases} \tag{2.2}$$

훈련데이터 $\{(\boldsymbol{x}_i, z_i^j)\}_{i=1}^n$에 대하여 지지벡터기계를 적합하여 얻은 분류함수를 f_j로 나타내면, 새로운 데이터 $\boldsymbol{x}$에 대한 클래스 예측값은 다음과 같다. 모든 j에 대하여 $f_j(\boldsymbol{x}) < 0$인 경우에는 $y = 1$로 예측하고, $f_j(\boldsymbol{x}) > 0$인 최소의 j가 존재하면 $y = j + 1$로 예측한다.

OVNF

$j\ (=1,\cdots,J-1)$에 대하여 새로운 출력변수의 값을 다음과 같이 정의하자. $i=1,\ldots,n$에 대하여

$$z_i^j=\begin{cases}1, & y_i=j+1\\ -1, & y_i=j.\end{cases} \tag{2.3}$$

훈련데이터 $\{(\boldsymbol{x}_i,z_i^j)\}_{i=1}^n$에 대하여 지지벡터기계를 적합하여 얻은 분류함수를 f_j로 나타내자. 새로운 데이터 $\boldsymbol{x}$에 대한 클래스 예측값은 다음과 같다. $j=1$에서부터 순차적으로 $J-1$까지 증가시켜 가며, $f_j(\boldsymbol{x})<0$이면 $y=j$로 예측하고 그렇지 않으면 클래스 $j+1$과 $j+2$를 비교하는 과정을 계속한다 (Kim and Ahn, 2010).

2.3. 비례오즈 모형

비례오즈 모형은 출력변수가 순서형일 때 적용 가능한 다범주 로지스틱 회귀모형이다. 순서형 출력변수 Y가 $j\ (=1,\ldots,J-1)$이하의 범주에 속할 누적 확률 $\mathbb{P}(Y\le j|\mathbf{x})$에 대한 누적 로짓 (cumulative logit)은

$$\text{logit}[\mathbb{P}(Y\le j|\mathbf{x})]=\log\frac{\mathbb{P}(Y\le j|\mathbf{x})}{1-\mathbb{P}(Y\le j|\mathbf{x})}$$

로 정의되며, 이를 이용한 비례오즈 모형은

$$\text{logit}[\mathbb{P}(Y\le j|\mathbf{x})]=\alpha_j+\beta^T\mathbf{x} \tag{2.4}$$

이다. 모형 (2.4)에서 j번째 누적 로짓은 절편 α_j을 가지고 있으며 α_j들은 j가 커지면 증가한다. 모든 로짓에 설명 변수가 미치는 효과는 $\boldsymbol{\beta}$로 동일하기 때문에 절편항 α_j의 값에 따라서 각 범주에 대한 예측확률이 정해진다 (Agresti, 2002).

3. 데이터 분석

이 절에서는 모의실험과 실제 온라인 리뷰 데이터에 대하여 2절에서 논의한 다범주 분류방법들의 예측 성능을 비교한다. 모든 데이터 분석은 R을 이용하였다.

3.1. 모의실험

우선 모의실험 데이터 생성모형을 설명하면 다음과 같다. 출력변수 Y는 $\{1,2,3\}$상의 균일분포를 따르고, 50개의 입력변수들 중 처음 5개의 변수 $X_1,\cdots,X_5$, 가운데 40개의 변수 $X_6,\cdots,X_{45}$, 마지막 5개의 변수 $X_{46},\cdots,X_{50}$들은 각각 긍정적인 단어, 중립적인 단어, 부정적인 단어들의 그룹을 나타낸다. 클래스 값이 $j\ (=1,2,3)$일 때 입력변수들의 세 그룹에 대한 포아송 모수는 $\lambda_{j1},\lambda_{j2},\lambda_{j3}$ 로 나타낸다. Table 3.1의 포아송 모수값에 따라 세 가지 시나리오를 상정하였으며, 클래스간에 포아송 모수 값의 차이에 대한 측도로

$$d=\sum_{i<j}\|p_i-p_j\|$$

를 이용한다. 여기서 $p_j=(\lambda_{j1},\lambda_{j2},\lambda_{j3})^T$이고 d값이 커질수록 클래스들간의 구분이 잘 된다고 볼 수 있다.

Table 3.1 Poisson parameters for 3 scenarios in simulation with $J = 3$

Scenario	d	Class	$X_1, \cdots, X_5$ (Positive words)	$X_6, \cdots, X_{45}$ (Neutral words)	$X_{46}, \cdots, X_{50}$ (Negative words)
A	1.131	1	0.1	0.1	0.5
		2	0.3	0.1	0.3
		3	0.5	0.1	0.1
B	1.366	1	0.1	0.1	0.5
		2	0.1	0.1	0.1
		3	0.5	0.1	0.1
C	1.624	1	0	0.1	0.5
		2	0.1	0.3	0.1
		3	0.5	0.1	0.0

본 연구에서는 다범주 분류문제에서 정분류율보다 더 나은 예측력 평가측도로 알려진 Hand와 Till (2001)의 M-AUC (multi-class area under the curve)를 평가기준으로 사용한다. M-AUC를 정의하기 위해 몇 가지 기호를 도입하면 다음과 같다. $\hat{A}(i|j)$는 랜덤하게 선택된 j번째 클래스의 데이터에 대하여 랜덤하게 선택된 i번째 클래스의 데이터보다 더 클래스 i에 속할 확률의 추정치가 낮을 확률을 의미한다. $\hat{A}(j|i)$는 $\hat{A}(i|j)$의 반대의 경우에 대한 확률을 나타낸다고 하자. M-AUC는

$$\frac{2}{J(J-1)} \sum_{i<j} \hat{A}(i,j)$$

로 정의된다. 여기서 $\hat{A}(i,j) = \frac{\hat{A}(i|j)+\hat{A}(j|i)}{2}$이다.

모의실험은 각 시나리오별로 다음과 같이 실시하였다. 앞에서 설명한 데이터 생성 방법에 따라 5,000개의 데이터를 생성하여 1,000개는 훈련데이터로 사용하고 나머지는 시험데이터로 사용하였다. 각 방법들의 객관적인 비교를 위하여 데이터 생성 및 분할, 모형 적합 및 예측의 전 과정을 100회 반복하여 M-AUC의 평균과 표준오차를 구하였다. Table 3.2는 모의실험 결과를 보여준다.

Table 3.2 Results from simulations with $J = 3$

Method	Ratio	M-AUC		
		Scenario A	Scenario B	Scenario C
OVNF	1:1:1	0.7795 (0.0006)	0.7738 (0.0021)	0.8437 (0.0030)
OVN		0.7881 (0.0006)	0.8265 (0.0006)	0.9411 (0.0004)
OVR		0.7616 (0.0005)	0.8200 (0.0006)	0.9389 (0.0004)
OVO		0.7801 (0.0006)	0.8180 (0.0006)	0.9334 (0.0004)
Odds		0.7946 (0.0005)	0.8254 (0.0005)	0.8859 (0.0004)
OVNF	1:2:2	0.7733 (0.0007)	0.7923 (0.0017)	0.8589 (0.0031)
OVN		0.7718 (0.0007)	0.8177 (0.0007)	0.9362 (0.0005)
OVR		0.7628 (0.0008)	0.8147 (0.0007)	0.9386 (0.0005)
OVO		0.7734 (0.0007)	0.8127 (0.0007)	0.9326 (0.0005)
Odds		0.7794 (0.0007)	0.8204 (0.0006)	0.8917 (0.0005)
OVNF	1:1:4	0.7394 (0.0010)	0.7317 (0.0027)	0.7904 (0.0043)
OVN		0.7650 (0.0008)	0.8016 (0.0008)	0.9240 (0.0007)
OVR		0.7215 (0.0007)	0.7735 (0.0017)	0.9301 (0.0006)
OVO		0.7401 (0.0009)	0.7825 (0.0010)	0.9271 (0.0006)
Odds		0.7487 (0.0011)	0.7787 (0.0010)	0.8672 (0.0006)

Table 3.2로부터 모의실험 결과를 해석해 보면 다음과 같다. d값이 커질수록 각 방법간의 예측력 차이가 뚜렷해지며, 클래스간의 비율에서 불균형 (imbalance)이 심해질수록 방법들은 성능이 전반적으로 떨어지는 패턴을 보인다. d값이 작은 경우 (시나리오 A)의 경우에는 클래스 비율이 1:1:1 혹은 1:2:2일

때 비례오즈 모형은 다른 방법에 비하여 성능이 상대적으로 좋게 나타났다. 그러나 d값이 큰 경우에는 순서성을 고려하는 OVN이 클래스 비율과 상관없이 제일 좋은 성능을 보인 반면 비례오즈 모형의 성능은 그리 좋지 않았다. 아마도 비례오즈 모형의 경우 범주별로 절편값만 다르기 때문에 다양한 분류경계를 주지 못하여 d가 커질수록 상대적인 성능이 떨어지는 것이 아닌지 의심된다.

이번에는 범주의 갯수를 더 늘려 $J = 5$인 모의실험을 추가로 실시하였다. 앞에서 실시한 $J = 3$인 경우와 마찬가지로 출력변수 Y는 $\{1, 2, 3, 4, 5\}$상의 균일분포를 따르고, 50개의 입력변수들 중 처음 5개의 변수 $X_1, \cdots, X_5$, 가운데 40개의 변수 $X_6, \cdots, X_{45}$, 마지막 5개의 변수 $X_{46}, \cdots, X_{50}$들은 각각 긍정적인 단어, 중립적인 단어, 부정적인 단어들의 그룹을 나타낸다. 클래스 값이 j $(= 1, \ldots, 5)$일 때 입력변수들의 세 그룹에 대한 포아송 모수는 $\lambda_{j1}, \ldots, \lambda_{j5}$ 로 나타낸다. Table 3.3의 포아송 모수값에 따라 세 가지 시나리오를 상정하였다.

Table 3.3 Poisson parameters for 3 scenarios in simulation with $J = 5$

Scenario	d	Class	$X_1, \cdots, X_5$ (Positive words)	$X_6, \cdots, X_{45}$ (Neutral words)	$X_{46}, \cdots, X_{50}$ (Negative words)
A	3.203	1	0.1	0.1	0.5
		2	0.1	0.1	0.4
		3	0.3	0.1	0.3
		4	0.4	0.1	0.1
		5	0.5	0.1	0.1
B	3.415	1	0.1	0.1	0.5
		2	0.1	0.1	0.3
		3	0.1	0.2	0.1
		4	0.3	0.1	0.1
		5	0.5	0.1	0.1
C	3.863	1	0	0.1	0.5
		2	0.1	0.2	0.3
		3	0.1	0.3	0.1
		4	0.3	0.2	0.1
		5	0.5	0.1	0.0

Table 3.4는 모의실험의 결과를 요약한다. 전반적인 패턴은 $J = 3$인 경우와 매우 유사하다. 다만 $J = 5$인 경우 다른 방법들에 대한 OVN의 상대적 성능이 $J = 3$의 경우에 비해 더 두드러지게 나타난다.

Table 3.4 Results from simulations with $J = 5$

Method	Ratio	M-AUC		
		Scenario A	Scenario B	Scenario C
OVNF	1:1:1	0.6286 (0.0052)	0.6390 (0.0028)	0.7258 (0.0006)
OVN		0.7570 (0.0006)	0.7862 (0.0005)	0.8608 (0.0004)
OVR		0.7335 (0.0007)	0.7565 (0.0006)	0.8207 (0.0005)
OVO		0.7396 (0.0006)	0.7619 (0.0006)	0.8426 (0.0005)
Odds		0.7736 (0.0005)	0.7875 (0.0005)	0.8346 (0.0004)
OVNF	1:2:2	0.6169 (0.0051)	0.6194 (0.0028)	0.7250 (0.0013)
OVN		0.7497 (0.0006)	0.7784 (0.0007)	0.8539 (0.0005)
OVR		0.7327 (0.0008)	0.7505 (0.0007)	0.8185 (0.0006)
OVO		0.7377 (0.0007)	0.7582 (0.0007)	0.8392 (0.0006)
Odds		0.7521 (0.0006)	0.7686 (0.0006)	0.8153 (0.0005)
OVNF	1:1:4	0.5834 (0.0056)	0.5978 (0.0036)	0.7033 (0.0010)
OVN		0.7428 (0.0008)	0.7559 (0.0008)	0.8372 (0.0007)
OVR		0.7072 (0.0013)	0.7312 (0.0013)	0.8067 (0.0007)
OVO		0.7222 (0.0009)	0.7354 (0.0012)	0.8201 (0.0009)
Odds		0.7179 (0.0007)	0.7239 (0.0008)	0.7701 (0.0006)

3.2. 온라인 리뷰 데이터

현재 운영되고 있는 웹 사이트로부터 영화, 앱게임 등에 관한 리뷰와 평점을 웹 스크랩 (web scraping) 기법으로 가져와서 2절에 설명된 텍스트에 대한 전처리를 하였다. R을 이용하여 웹 데이터를 스크랩하는 기법은 Munzert 등 (2015)을 참고하기 바란다.

영화 리뷰 데이터

영화 리뷰 사이트 `http://www.imdb.com`에서는 이용자들로부터 얻은 영화들에 대한 10점 척도의 평점과 리뷰를 제공한다. 이 웹사이트에서 2015년 1월부터 2015년 6월까지 개봉된 영화에 대한 리뷰들을 수집하였다. 평점이 1점에서 3점인 경우를 클래스 1, 4점에서 7점인 경우를 클래스 2, 8점에서 10점을 클래스 3으로 하였고, 각 클래스의 주변확률에 비례하여 총 7,000개의 리뷰들을 가져와 전처리 하였다.

앱게임 리뷰 데이터

아마존 (`http://www.amazon.com`)에서는 이용자들이 구매한 앱게임에 대한 5점 척도의 평점 및 리뷰를 제공한다. 적절성 (relevance), 최신순 (new release)의 정렬 기준에 따라 나오는 143개의 앱게임에서 상위 10개의 리뷰를 각 평점별로 스크랩하여 총 7,150개의 리뷰 데이터를 얻었다.

랜덤분할에 의하여 방법론들의 비교하기 전에 우선 평점과 연관성이 높은 단어들을 Figure 3.1과 같이 워드클라우드로 나타내어 보았다. 카이제곱 검정의 p-값이 0.001이하인 단어들만 선택하였다. Figure 3.1 (a)는 영화 리뷰에 대한 워드클라우드로 waste, worst와 같은 부정적인 단어들과 영화에 대한 느낌을 표현하는 단어인 perfect, amazing, boring, terrible과 같은 감정을 나타내는 형용사들이 출현빈도가 많고 평점과 높은 상관관계를 갖음을 알 수 있다. Figure 3.1 (b)와 (c)는 각각 적절성, 최신순으로 선택된 리뷰에 대한 워드클라우드를 보여준다. 두 워드클라우드를 비교해 보면 최신 순으로 선택된 게임 리뷰에서 추출된 단어들의 갯수가 상대적으로 적다는 것을 확인할 수 있다. 이는 최신 앱게임에 대한 리뷰 문서의 분량이 적절성순으로 정렬한 리뷰들에 비해서 상대적으로 적음을 의미한다. 앱게임의 경우 영화 리뷰에서와는 다르게 게임의 설치와 운영 등과 관련된 단어들인 deleted, update과 같은 단어들이 평점과 높은 상관관계를 갖음을 알 수 있다.

Figure 3.1 Word clouds for movie and game review data.

이제 데이터에 대한 랜덤분할을 통하여 여러 가지 지지벡터기계와 비례오즈 모형의 성능을 비교하기로 한다. 매 분할에서는 전체 리뷰 데이터를 훈련과 시험데이터로 1:1로 랜덤하게 분할하였다. 훈련데이터를 이용하여 카이제곱 검정의 p-값 0.001이하인 단어들을 추출하고, 추출된 단어들이 리뷰에 나오는 횟수와 평점을 각각 입력 및 출력변수로 놓고 다범주 분류방법들을 적용하였다. 시험데이터에서는 각 분류방법에 대하여 M-AUC값을 구하였다. 이러한 전 과정을 100회 반복하여 Table 3.5과 Table 3.6와 같이 M-AUC의 평균과 표준오차를 구하였다.

Table 3.5 Results from movie review data

Method	M-AUC	# of words
OVNF	0.751 (.001)	324.85 (1.57)
OVN	0.763 (.001)	325.62 (1.45)
OVR	0.776 (.001)	322.07 (1.45)
OVO	0.771 (.001)	326.58 (1.44)
Odds	0.776 (.001)	324.55 (1.56)

Table 3.5과 Table 3.6는 각각 영화 리뷰와 앱게임 리뷰 데이터에 대한 결과를 보여준다. 카이제곱 통계량에 의해 선택된 단어의 갯수는 영화 리뷰에서는 대략 320개, 앱게임 리뷰에서는 추천도순의 경우 대략 100개, 최신순의 경우 대략 50개이다. 앱게임 리뷰는 스마트폰에서 작성하는 경우가 많기 때문에 영화리뷰에 비해 길이가 상대적으로 짧고, 최신 앱게임의 경우 기존의 앱게임에 비해 리뷰가 상대적으로 적기 때문일 것으로 추측된다. 영화 리뷰의 경우 M-AUC값 기준으로 OVR과 비례오즈 모형, OVO, ONV과 OVNF 순으로 나타났다. 앱게임 리뷰의 경우 M-AUC는 비례오즈 모형, OVN, OVO, OVR, OVNF 순으로 나타났다. 두 데이터 모두에서 비례오즈 모형의 성능이 제일 좋은 것으로 볼 때, 클래스 간의 분류경계가 절편차이로 평행한 직선으로 비교적 잘 표현된다고 볼 수 있다. 또한 모의실험과 유사하게 OVO도 비교적 안정적인 성능을 보인다. 게임 리뷰의 경우를 제외하고 순서성을 고려한 지지벡터기계의 방법인 OVN이나 OVNF의 성능이 그리 좋지 않음을 알 수 있다.

Table 3.6 Results from app and games review data

Method	Relevance		New release	
	M-AUC	# of words	M-AUC	# of words
OVNF	0.632 (.002)	105.29 (.66)	0.639 (.003)	56.43 (.46)
OVN	0.714 (.000)	104.63 (.73)	0.708 (.001)	57.34 (.43)
OVR	0.698 (.001)	104.66 (.71)	0.672 (.002)	56.83 (.45)
OVO	0.708 (.001)	102.92 (.64)	0.694 (.001)	56.65 (.46)
Odds	0.721 (.000)	104.72 (.67)	0.726 (.000)	56.98 (.45)

4. 결론

본 논문에서는 카이제곱 통계량을 이용하여 온라인 리뷰 데이터에서 평점과 연관성이 높은 단어들을 추출한 후 각 리뷰에서 추출된 단어들의 출현빈도를 이용하여 평점을 예측하는 다범주 분류문제를 고려하였다. 우선 분석에 사용된 데이터의 특징을 살펴보면, 긍정과 부정으로 이루어진 이진 분류 보다는 범주의 갯수가 3개 정도의 다범주 분류가 적합한 것으로 보이며, 특정 단어들이 나오는지 여부가 클래스와 연관성이 매우 큰 것으로 보인다. 이범주 지지벡터기계에 기반한 여러 가지 다범주 지지벡터기계와 비례오즈 모형의 성능을 모의실험 및 실제 영화와 게임 리뷰 데이터 분석을 통하여 비교하였다. 모의실험에서는 OVN과 OVR의 성능이 전반적으로 높게 나오는 반면 실제 데이터에서는 비례오즈 모형의 성능이 높게 나온 것으로 보아 예측성능이 데이터의 특성에 따라 달라짐을 알 수 있다. 추가적인 모의실험과 실제 데이터의 분석을 통하여 더 연구해 볼 필요가 있을 것으로 생각된다.

본 연구에 대한 후속 연구 방향으로는 다음과 같은 것들을 생각해 볼 수 있다. 첫째, 비례오즈 모형에 대하여 스플라인 기저 등을 이용한 일반화 가법모형 (generalized additive model)을 고려할 수 있다. 순서성을 반영해야 하므로 추정된 분류경계들간에 엇갈림이 없는 조건 (non-crossing constraint)하에서 추정해야 할 것이다. 비선형 비례오즈 모형은 더 다양한 분류경계를 모형화 할 수 있기 때문에 예측력을 더욱 높여줄 수 있을 것으로 기대된다. 둘째, 의사결정 보류 옵션 (reject option)을 가지는 지지벡터기계나 순위 모형 (ranking model)에 기반한 방법 등도 고려해 볼 수 있을 것이다.

References

Agresti, A. (2002). *Categorical data analysis*, 2nd Ed., Wiley, New Jersey

Bae, K. Y., Park, J.-H., Kim, J. S., and Chae, M., Kang, M., and Lee, Y.-S. (2013). Analysis of the abstracts of research articles in food related to climate change using a text-mining algorithm. *Journal of the Korean Data & Information Science Society*, **24**, 1429-1437.

Chae, M., Kang, M., and Kim, Y. (2013). Documents recommendation using large citation data. *Journal of the Korean Data & Information Science Society*, **24**, 999-1011.

Hand, D. J. and Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, **45**, 171-186.

Hsu, C.-W. and Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines, *IEEE Transactions on neural networks*, **13**, 415-425.

Kim, K.-J. and Ahn, H.C. (2010). Customer level classification model usings ordinal multiclass support vector machines. *Asia Pacific Journal of Information Systems*, **20**, 23-37.

Kim, S. O., Lee, S. Y., Lee, S. J., and Lee, H. C. (2013). A study of development for movie recommendation system algorithm using filtering. *Journal of the Korean Data & Information Science Society*, **24**, 803-813.

Kim, S. and Kim, N. (2014). A Study on the effect of using sentiment lexicon in opinion classification. *Journal of Intelligence and Information Systems*, **20**, 133-148.

Lee, H and Hong, T. (2015). Terms based sentiment classification for online review using support vector machine. *Information Systems Review*, **17**, 49-64.

Lee, H. and Suh, Y. (2014). Social media comparative analysis based on multidimensional scaling. *Journal of the Korean Data & Information Science Society*, **25**, 665-676.

Liu, B. (2012). *Sentiment analysis and opinion mining*, Morgan & Claypool Publishers, San Bernardino, California.

Munzert, S., Rubba, C., Meißner, P. and Nyhuis, D. (2015). *Automated data collecction with R*, Wiley, West Sussex, United Kingdom.

Vapnik, V. (1995). *The nature of statistical learning*, Springer, New York.

Journal of the Korean Data & Information Science Society
2016, **27**(4), 845–854

http://dx.doi.org/10.7465/jkdi.2016.27.4.845
한국데이터정보과학회지

Classification of ratings in online reviews[†]

Dongjun Choi[1] · Hosik Choi[2] · Changyi Park[3]

[13]Department of Statistics, University of Seoul
[2]Applied Information Statistics, Kyonggi University

Received 29 June 2016, revised 18 July 2016, accepted 22 July 2016

Abstract

Sentiment analysis or opinion mining is a technique of text mining employed to identify subjective information or opinions of an individual from documents in blogs, reviews, articles, or social networks. In the literature, only a problem of binary classification of ratings based on review texts in an online review. However, because there can be positive or negative reviews as well as neutral reviews, a multi-class classification will be more appropriate than the binary classification. To this end, we consider the multi-class classification of ratings based on review texts. In the preprocessing stage, we extract words related with ratings using chi-square statistic. Then the extracted words are used as input variables to multi-class classifiers such as support vector machines and proportional odds model to compare their predictive performances.

Keywords: Multi-class classification, opinion mining, sentiment analysis, word cloud.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2015R1D1A1A01059984).
[1] Master student, Department of Statistics, University of Seoul, Seoul 02504, Korea.
[2] Assistant professor, Department of Applied Information Statistics, Kyonggi University, Suwon 16227, Korea.
[3] Corresponding author: Associate professor, Department of Statistics, University of Seoul, Seoul 02504, Korea. E-mail: park463@uos.ac.kr