The Korean Journal of Applied Statistics
2022, Vol. 35, No. 4, 469–484

DOI: https://doi.org/10.5351/KJAS.2022.35.4.469
Print ISSN1225-066X / Online ISSN 2383-5818

Properties of chi-square statistic and information gain for feature selection of imbalanced text data

Hye In Mun[a], Won Son[1,a]

[a]Department of Statistics, Dankook University

Abstract

Since a large text corpus contains hundred-thousand unique words, text data is one of the typical large-dimensional data. Therefore, various feature selection methods have been proposed for dimension reduction. Feature selection methods can improve the prediction accuracy. In addition, with reduced data size, computational efficiency also can be achieved. The chi-square statistic and the information gain are two of the most popular measures for identifying interesting terms from text data. In this paper, we investigate the theoretical properties of the chi-square statistic and the information gain. We show that the two filtering metrics share theoretical properties such as non-negativity and convexity. However, they are different from each other in the sense that the information gain is prone to select more negative features than the chi-square statistic in imbalanced text data.

Keywords: chi-square statistic, feature selection, imbalanced data, information gain, text data

1. 서론

텍스트 데이터의 분석을 위해 각 문서에 포함된 단어의 출현빈도를 기록한 문서-단어행렬(document-term matrix; DTM)이 자주 사용된다. 한편, 다양한 범주의 많은 문서들을 모아둔 텍스트 데이터의 경우 수만에서 수십만 개의 서로 다른 단어들이 포함되어 있는 경우가 관찰되기도 한다. 문서-단어행렬을 작성하기 위한 전처리(preprocessing) 과정에서 불용어(stop-words)와 출현빈도가 낮은 단어들을 제거하는 과정을 거치지만 여전히 필요에 비해 많은 단어들이 문서-단어행렬에 포함되어 있는 경우가 흔하다. 이렇게 불필요한 많은 단어들이 포함되어 있는 경우 문서-단어행렬의 차원이 커지게 되는데, 이러한 고차원 데이터의 경우 분석의 정확성과 계산의 효율성에서 문제가 생기기도 하는 것으로 알려져 있다.

이와 같은 문제점을 해결하기 위해 텍스트 데이터의 특성을 파악하는 데 유용한 단어들만 선택하는 특징 선택(feature selection) 방법들이 자주 활용되고 있다. 텍스트 데이터에서 유용한 단어를 선택하기 위한 대표적인 지표로는 카이제곱통계량(chi-square statistic), 오즈비(odds ratio), 정보이득(information gain) 등을 들 수 있다. 이들 통계량은 단어와 문서의 범주에 대한 이차원분할표(2-way contingency table)의 정보를 요약하는 지표라는 공통점이 있다.

한편, 이들 통계량의 성능에 대한 판단은 연구 결과에 따라 차이가 있다. Yang과 Pedersen (1997)은 Reuters-22173 데이터와 OHSUMED 데이터를 이용하여 카이제곱통계량과 정보이득이 포함된 5개 지표의 성능을 비교하였는데 전반적으로 카이제곱통계량의 성능이 좋은 것으로 평가되었다. Mladenić과 Grobelnik (1999)은 야후의 웹 분류체계를 이용하여 성능을 비교하였는데 정보이득과 오즈비가 포함된 11개의 지표 중 오즈비가

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

Published 31 August 2022 / journal homepage: http://kjas.or.kr

Table 1: 2-way contingency table for the term w and class c

		Class membership		Total
		c	not c ($\bar{c}$)	
Existence of the term w	Present	n_{11}	n_{01}	$n_{\cdot 1} = n_{11} + n_{01}$
	Absent	n_{10}	n_{00}	$n_{\cdot 0} = n_{10} + n_{00}$
Total		$n_{1\cdot} = n_{11} + n_{10}$	$n_{0\cdot} = n_{01} + n_{00}$	$n_{\cdot\cdot} = N$

상대적으로 좋은 성능을 보인 반면, 정보이득의 성능은 좋지 않은 것으로 나타났다. Forman (2003)은 카이제곱통계량과 정보이득, 오즈비를 포함한 12개의 단어 선택 지표들을 Reuters-21578, OHSUMED 등 다양한 데이터에 적용하였는데 선택된 단어의 수가 적을 때 정보이득이 카이제곱통계량에 비해 항상 좋은 결과를 보이는 것으로 관찰되었다.

이렇게 실제 데이터를 이용하여 단어선택 지표의 성능을 비교하는 경우 연구자에 따라 서로 다른 결론이 도출될 수 있음을 확인할 수 있다. 따라서 각 지표들의 이론적 특징을 비교해 볼 필요가 있지만 이와 관련된 연구결과는 많지 않다. 단어 선택 지표들의 이론적 성질에 대한 연구로는 카이제곱통계량과 오즈비의 성질을 비교한 Son (2020)의 연구를 들 수 있다.

이 연구에서는 이차원분할표 기반 단어 선택 지표들 중 카이제곱통계량과 정보이득의 특징을 비교해보려 한다. 카이제곱통계량과 정보이득의 경우 각 집단들이 동질적이라는 귀무가설 아래에서 점근분포가 카이제곱분포를 따른다는 것이 알려져 있는 등 많은 유사성이 있다. 하지만, 앞에서 살펴본 바와 같이 실제 데이터 분석에서는 두 지표의 성능에 차이가 있으며 이런 점을 고려할 때 두 지표의 차이에 대한 면밀한 검토가 필요하다.

이 논문은 다음과 같이 구성된다. 2절에서는 오즈비, 카이제곱통계량, 정보이득 등 이차원분할표 기반의 단어 선택 기법들에 대해 소개한다. 3절에서는 카이제곱통계량과 정보이득의 비음성(non-negativity), 볼록성(convexity)과 불균형 텍스트 데이터에서의 관계 등에 대해 살펴본다. 4절에서는 실제 데이터를 이용하여 카이제곱통계량과 정보이득의 관계를 확인해보고 분류모형에서의 성능을 비교해본다. 마지막으로 결과를 정리하고 앞으로의 연구 방향을 소개한다.

2. 텍스트 데이터에 대한 필터링 방식의 단어 선택 방법

Table 1은 텍스트 데이터에 포함되어 있는 각 문서들이 특정 주제 c에 해당되는지, 또 특정 단어 w를 포함하고 있는지를 기준으로 문서들을 분류하여 문서의 빈도를 기록한 이차원분할표이다. 텍스트 데이터에 C개의 범주와 W개의 단어가 있는 경우 CW개의 이차원분할표가 작성된다. Table 1의 n_{ij}에서 문서가 범주 c에 포함되면 첫 번째 인덱스 i가 1, 그렇지 않으면 0이고, 문서가 단어 w를 포함하면 두 번째 인덱스 j가 1, 아니면 0으로 기록된다. 즉, n_{11}은 단어 w를 포함하면서 범주 c에 속하는 문서의 수를, n_{10}은 단어 w를 포함하지 않으면서 범주 c에 속하는 문서의 수를 나타낸다. 또, n_{01}은 단어 w를 포함하면서 범주 c에 속하지 않는 문서의 수를, n_{00}은 단어 w를 포함하지 않으면서 범주 c에 속하지 않는 문서의 수를 나타낸다. 관측값의 개수 $N = n_{11} + n_{10} + n_{01} + n_{00}$에 대해 각 셀의 비율을 $p_{ij} = n_{ij}/N$로 정의한다.

이렇게 작성된 이차원분할표의 정보는 텍스트 데이터 분류에 유용한 단어를 식별하기 위해 활용할 수 있다. 예를 들어 신문 기사를 주제별로 분류할 때 경제 관련 기사를 식별하기 위해 도움이 되는 "주가", "환율", "금리" 등의 단어는 경제 관련 기사에는 자주 등장하지만 경제와 관련 없는 기사에는 출현빈도가 낮을 것으로 예상할 수 있다. 따라서 특정 범주와 단어의 상관관계가 높을 때 Table 1에서 대각선 부분에 해당되는 셀의 숫자가 크고 이외의 셀의 숫자는 작아진다고 볼 수 있다. 카이제곱통계량, 오즈비, 정보이득 등은 모두 Table 1과 같은 이차원분할표의 정보를 요약하여 단어의 등장 여부와 문서의 범주 사이의 관계를 측정하는 지표이다.

2.1. 오즈비

오즈비는 범주 c에 속하는 문서 중 단어 w를 포함하고 있는 문서와 그렇지 않은 문서의 비 p_{11}/p_{10}을 범주 c에 속하지 않는 문서 중 단어 w를 포함하고 있는 문서와 그렇지 않은 문서의 비 p_{01}/p_{00}으로 나눈 값에 해당되며

$$\text{OR} = \frac{p_{11}/p_{10}}{p_{01}/p_{00}} = \frac{p_{11}p_{00}}{p_{10}p_{01}} = \frac{n_{11}n_{00}}{n_{10}n_{01}} \tag{2.1}$$

과 같이 정의할 수 있다. 따라서 이차원분할표에서 대각선 원소 n_{11}과 n_{00}의 값이 클수록 오즈비가 커지고 반대로 대각선에 해당되지 않는 원소 n_{10}과 n_{01}의 값이 클수록 오즈비가 0에 가까운 작은 값을 가진다.

2.2. 카이제곱통계량

카이제곱통계량은 이차원분할표의 각 셀별 기댓값과 실제 셀 값의 차이를 이용하여 계산할 수 있으며

$$\chi^2 = \sum_{i=0}^{1}\sum_{j=0}^{1}\frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{2.2}$$

와 같이 정의된다. 여기서 E_{ij}는 (i, j) 셀의 기댓값으로 $p_{i\cdot} = n_{i\cdot}/N = \sum_{j=0}^{1} n_{ij}/N$, $p_{\cdot j} = n_{\cdot j}/N = \sum_{i=0}^{1} n_{ij}/N$에 대해 $E_{ij} = Np_{i\cdot}p_{\cdot j}$로 정의된다.

식 (2.2)의 카이제곱통계량은

$$\chi^2 = \frac{N(n_{11}n_{00} - n_{10}n_{01})^2}{n_{1\cdot}n_{\cdot 1}n_{0\cdot}n_{\cdot 0}} = \frac{N(p_{11}p_{00} - p_{10}p_{01})^2}{p_{1\cdot}p_{0\cdot}p_{\cdot 1}p_{\cdot 0}} \tag{2.3}$$

과 같이 변환할 수 있으며 분자의 $p_{11}p_{00} - p_{10}p_{01}$은 오즈비의 분자에서 분모를 뺀 형태에 해당된다. 식 (2.3)에서 카이제곱통계량은 이차원분할표의 대각선 원소들의 곱이 비대각선 원소들의 곱에 비해 클 때 또는 비대각선 원소들의 곱이 대각선 원소들의 곱에 비해 클 때 큰 값을 가지고 대각선 원소들의 곱과 비대각선 원소들의 곱이 같게 되어 오즈비가 1일 때 최솟값 0을 가지게 됨을 알 수 있다. 즉, Son (2020)에 제시된 바와 같이 오즈비는 단어 w가 범주 c와 양의 연관성을 가질 때만 큰 값을 가지는 반면, 카이제곱통계량은 단어 w가 범주 c와 음의 연관성을 가질 때도 큰 값을 가진다는 점에서 차이가 있다. 예를 들어 경제 관련 주제를 다루고 있는 문서에는 "연주회", "피아노", "바이올린" 등의 단어가 등장하는 경우가 많지 않으므로 이 단어들이 포함된 문서는 경제 관련 주제가 아니라고 판단할 수 있다. 이렇게 단어 w가 범주 c와 양의 연관성을 가질 때 w를 양변수(positive feature)라 하고 음의 연관성을 가질 때 w를 음변수(negative feature)라 한다. 한편, 카이제곱통계량은 동질성이 성립한다는 귀무가설 아래에서 점근적으로 카이제곱분포를 따르므로 통계적 가설검정 절차를 이용하여 단어를 선택하는 방법을 생각해볼 수 있다. 하지만 실제 텍스트 데이터에 카이제곱통계량을 적용하여 가설검정을 진행하는 경우 지나치게 많은 단어가 선택되는 경향을 발견할 수 있다. 따라서 실제 데이터 분석에서는 카이제곱검정 절차를 사용하는 대신 카이제곱통계량을 기반으로 주관적인 판단하에 단어를 선택하는 경우가 많다 (Forman, 2003; Kou 등, 2020).

2.3. 정보이득

정보이득은 단어 w를 이용하였을 때 엔트로피(entropy)가 감소하는 정도를 측정하는 지표로

$$\text{IG}(c, w) = \text{H}(c) - \text{H}(c\,|w) \tag{2.4}$$

와 같이 표현할 수 있다. 여기서 $\text{H}(c) = -\sum_{i=0,1} p_{i\cdot}\log p_{i\cdot}$는 범주 c만 이용했을 때의 엔트로피를, $\text{H}(c\,|w) = -\sum_{i=0,1}\sum_{j=0,1} p_{ij}\log(p_{ij}/p_{\cdot j})$는 단어 w의 등장 여부에 대한 정보를 추가하였을 때의 엔트로피를 의미한다.

한편, $p_{i\cdot} = p_{i0} + p_{i1}$이므로 식 (2.4)는

$$\begin{aligned} \mathrm{IG}(c,w) &= -\sum_{i=0,1}\left(p_{i\cdot}\log p_{i\cdot} - \sum_{j=0,1} p_{ij}\log\frac{p_{ij}}{p_{\cdot j}}\right) \\ &= \sum_{i=0,1}\sum_{j=0,1} p_{ij}\log\frac{p_{ij}}{p_{i\cdot}p_{\cdot j}} \end{aligned} \tag{2.5}$$

로 나타낼 수도 있다. 식 (2.5)에서 정보이득은 이차원분할표에서의 가능도비 검정통계량에 해당되며 동질성에 대한 귀무가설 아래에서 점근적으로 카이제곱분포를 따름을 알 수 있다. 이하에서는 $\mathrm{IG}(c,w)$를 IG로 나타내기로 한다.

3. 카이제곱통계량과 정보이득의 성질

카이제곱통계량과 정보이득은 이차원분할표에 기반을 둔 통계지표이다. 특히 동질성에 대한 귀무가설 아래에서는 두 통계량이 점근적으로 카이제곱분포를 따른다는 사실이 알려져 있다. 두 지표의 이러한 유사성에도 불구하고 선행연구에서는 실제 데이터의 단어 선택에 적용되었을 때 그 결과에 차이가 있는 것으로 나타났다. 이 절에서는 두 지표의 공통점과 차이점을 확인해본다.

3.1. 두 지표의 공통점

먼저 두 지표는 항상 0 이상의 값을 가지는 비음성(non-negative)지표이다. 카이제곱통계량의 경우 식 (2.2)의 정의로부터 각 항들이 항상 0 이상이므로 음이 아님은 자명하다. 정보이득의 비음성은 옌센의 부등식(Jensen's inequality)을 이용하여 보일 수 있다. 옌센의 부등식은 볼록함수 $f(\cdot)$와 실수 $\lambda \in (0,1)$에 대해

$$\lambda f(x) + (1-\lambda)f(y) \geq f(\lambda x + (1-\lambda)y)$$

가 성립함을 의미한다. 함수 $-\log(\cdot)$는 볼록함수이므로 옌센의 부등식을 이용하면 $i = 0, 1$에 대해

$$\begin{aligned} \sum_{j=0,1} p_{ij}\log\frac{p_{ij}}{p_{i\cdot}p_{\cdot j}} &= p_{i1}\log\frac{p_{i1}}{p_{i\cdot}p_{\cdot 1}} + p_{i0}\log\frac{p_{i0}}{p_{i\cdot}p_{\cdot 0}} = -p_{i\cdot}\left(\frac{p_{i1}}{p_{i\cdot}}\log\frac{p_{i\cdot}p_{\cdot 1}}{p_{i1}} + \frac{p_{i0}}{p_{i\cdot}}\log\frac{p_{i\cdot}p_{\cdot 0}}{p_{i0}}\right) \\ &\geq -p_{i\cdot}\log\left(\frac{p_{i1}}{p_{i\cdot}}\frac{p_{i\cdot}p_{\cdot 1}}{p_{i1}} + \frac{p_{i0}}{p_{i\cdot}}\frac{p_{i\cdot}p_{\cdot 0}}{p_{i0}}\right) \\ &= -p_{i\cdot}\log(p_{\cdot 1} + p_{\cdot 0}) = 0 \end{aligned}$$

이고 정보이득은 0 이상의 값을 가짐을 알 수 있다.

이제 두 지표의 최솟값을 생각해보자. 카이제곱통계량의 경우 변형된 식 (2.3)을 통해

$$p_{11}p_{00} - p_{10}p_{01} = 0 \tag{3.1}$$

일 때 카이제곱통계량의 값이 0이 되고 카이제곱통계량의 비음성에서 이 때 카이제곱통계량이 최소가 되는 것을 알 수 있다. 한편, 정보이득의 각 항은 $\bar{i} = 1 - i$, $\bar{j} = 1 - j$로 정의할 때

$$\begin{aligned} p_{ij}\log\frac{p_{ij}}{p_{i\cdot}p_{\cdot j}} &= -p_{ij}\log\frac{\left(p_{ij} + p_{i\bar{j}}\right)\left(p_{\bar{i}j} + p_{ij}\right)}{p_{ij}} = -p_{ij}\log\left\{\left(p_{ij} + p_{i\bar{j}} + p_{\bar{i}j}\right) + \frac{p_{i\bar{j}}p_{\bar{i}j}}{p_{ij}}\right\} \\ &= -p_{ij}\log\left(1 - p_{\bar{i}\,\bar{j}} + p_{\bar{i}\,\bar{j}}\frac{p_{i\bar{j}}p_{\bar{i}j}}{p_{ij}p_{\bar{i}\,\bar{j}}}\right) \end{aligned}$$

와 같이 표현할 수 있다. 이 식에서 $p_{i\bar{j}}p_{\bar{i}j}/(p_{ij}p_{\bar{i}\,\bar{j}})$는 오즈비의 역수에 해당되므로 정보이득의 각 항은 오즈비가 1일 때 0이 됨을 알 수 있다.

다음으로 범주 i가 주어지면 $p_{1\cdot}$과 $p_{0\cdot}$도 결정되고 이 때 카이제곱통계량은 p_{11}의 볼록(convex)함수임을 보일 수 있다. 마찬가지로 정보이득도 p_{11}의 볼록함수이다. 카이제곱통계량과 정보이득의 볼록성에 대한 증명은 부록에서 다루기로 한다. 더 일반적으로 카이제곱통계량과 정보이득은 p_{11}뿐만 아니라 각 변수 p_{10}, p_{01}, p_{00}에 대해서도 볼록함수에 해당된다.

3.2. 불균형 데이터에서의 두 지표의 차이점

이상에서 살펴본 바와 같이 카이제곱통계량과 정보이득은 많은 공통점을 가지고 있지만 두 지표의 차이점도 존재한다. Table 1과 같은 형태로 주어진 가상의 텍스트 데이터를 생각해보자. 표에 제시된 데이터에서 $n_{1\cdot} = n_{11} + n_{10}$건의 문서는 범주 c에 속하고 나머지 $n_{0\cdot}$건의 문서는 범주 c에 속하지 않는다. 또 $n_{\cdot 1} = n_{11} + n_{01}$건의 문서에는 단어 w가 포함되어 있고 나머지 $n_{\cdot 0}$건의 문서에는 단어 w가 포함되어 있지 않다. Table 2의 예에서 살펴볼 수 있듯이 여러 범주로 구성된 데이터의 경우 한 범주가 차지하는 비중이 크지 않으므로 특정한 범주 c에 속하는 문서에 비해 c에 속하지 않는 문서가 훨씬 많은 경우, 즉 $n_{0\cdot} \gg n_{1\cdot}$인 불균형 데이터를 자주 관찰할 수 있다. 또, 관사, 대명사, 조동사 등의 불용어를 제외하면 특정한 단어 w가 많은 문서에 등장하는 경우는 흔하지 않으므로 n_{00}의 값이 다른 셀의 값보다 커서 $n_{00} > n_{10}, n_{01}, n_{11}$인 이차원분할표가 자주 관찰된다.

한편, 앞에서 살펴본 바와 같이 단어 w가 포함된 문서들을 범주 c에 속하는 문서로, w가 포함되지 않은 문서들을 범주 c에 속하지 않는 문서로 판단할 수 있을 때 w를 양변수라 한다. 이와 반대로 w가 포함된 문서들을 범주 c에 속하지 않는 문서로, w가 포함되지 않은 문서들을 범주 c에 속하는 문서로 판단할 수 있을 때 w를 음변수라 한다. 보다 구체적으로 단어 w와 범주 c의 오즈비가 1보다 클 때, 즉 $n_{11}n_{00} > n_{10}n_{01}$일 때, w가 c의 양변수이고, w와 c의 오즈비가 1보다 작을 때, 즉 $n_{11}n_{00} < n_{10}n_{01}$이 성립할 때 w를 c의 음변수로 정의할 수 있다. 특히 불균형 텍스트 데이터에서는 음변수에 대해서는 p_{11} 값이 0에 가까운 경우가 자주 관찰되고 반대로 양변수에 대해서는 p_{10} 또는 p_{01} 값이 0에 가까운 경우가 자주 관찰된다.

3.2.1. $p_{11} \approx 0$이고 $p_{11}p_{00} \ll p_{10}p_{01}$인 음변수일 때

이제 $n_{11}n_{00} \ll n_{10}n_{01}$이고 $p_{11} = n_{11}/N \approx 0$인 음변수에 대해 카이제곱통계량과 정보이득의 관계를 살펴보자. $n_{00} > n_{10}, n_{01}, n_{11}$을 가정하였으므로 부등식 $n_{11}n_{00} \ll n_{10}n_{01}$이 성립하려면 $n_{11} \ll n_{10}, n_{01}$이라 할 수 있다. 따라서 식 (2.3)의 카이제곱통계량은

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

으로 근사된다. 또 같은 가정 아래에서 로그함수의 테일러 전개를 이용하면

$$\begin{aligned}
p_{11}\log\frac{p_{11}}{p_{1\cdot}p_{\cdot 1}} &\approx p_{11}\log\frac{p_{11}}{p_{10}p_{01}} \approx 0,\\
p_{10}\log\frac{p_{10}}{p_{1\cdot}p_{\cdot 0}} &\approx p_{10}\log\frac{p_{10}}{p_{10}(p_{10}+p_{00})} \approx -p_{10}\log(1-p_{01}) \approx p_{10}p_{01},\\
p_{01}\log\frac{p_{01}}{p_{0\cdot}p_{\cdot 1}} &\approx p_{01}\log\frac{p_{01}}{(p_{01}+p_{00})p_{01}} \approx -p_{01}\log(1-p_{10}) \approx p_{10}p_{01},\\
p_{00}\log\frac{p_{00}}{p_{0\cdot}p_{\cdot 0}} &\approx p_{00}\log\frac{1-p_{01}-p_{10}}{(1-p_{10})(1-p_{01})} \approx p_{00}\log\left(1-\frac{p_{10}p_{01}}{(1-p_{10})(1-p_{01})}\right) \approx \frac{-p_{00}p_{10}p_{01}}{(1-p_{10})(1-p_{01})}
\end{aligned}$$

이고 정보이득은 근사적으로

$$\mathrm{IG} \approx p_{10}p_{01}\left(2 - \frac{p_{00}}{(1-p_{10})(1-p_{01})}\right) \tag{3.3}$$

로 표현된다. 따라서 식 (3.2)와 (3.3)에서 카이제곱통계량과 정보이득 사이에는

$$\chi^2 \approx N\frac{n_{01}}{n_{01}+n_{00}}\frac{n_{10}}{n_{10}+n_{00}} = \frac{Np_{10}p_{01}}{(p_{01}+p_{00})(p_{10}+p_{00})} \approx \frac{N}{p_{0\cdot}p_{\cdot 0}\left(2-\frac{p_{00}}{(1-p_{10})(1-p_{01})}\right)}\mathrm{IG}$$

와 같은 관계식이 성립함을 알 수 있다. 이 식에서 분모의 일부분인

$$2-\frac{p_{00}}{(1-p_{10})(1-p_{01})} = 1+\frac{p_{11}+p_{10}p_{01}}{(1-p_{10})(1-p_{01})} \approx 1+\frac{p_{10}p_{01}}{(1-p_{10})(1-p_{01})}$$

은 $p_{10}+p_{01} \ll 1$일 때, 즉 $p_{00} \approx 1$일 때 1에 가까운 값을 가지므로 카이제곱통계량과 정보이득 사이에는 근사관계식

$$\chi^2 \approx N\,\mathrm{IG} \tag{3.4}$$

이 성립한다.

3.2.2. $p_{01} \approx 0$이고 $p_{10}p_{01} \ll p_{11}p_{00}$인 양변수일 때

다음으로 $n_{10}n_{01} \ll n_{11}n_{00}$이고 $p_{01} = n_{01}/N \approx 0$인 양변수에 대해 카이제곱통계량과 정보이득의 관계를 살펴보자. $n_{00} > n_{10}, n_{01}, n_{11}$을 가정하였으므로 $n_{10}n_{01} \ll n_{11}n_{00}$인 양변수일 때는 $n_{11} \approx 0$일 수도 있다. 따라서 식 (2.3)의 카이제곱통계량은

$$\chi^2 = \frac{N(n_{11}n_{00}-n_{10}n_{01})^2}{n_{1\cdot}n_{0\cdot}n_{\cdot 1}n_{\cdot 0}} \approx \frac{N(n_{11}n_{00})^2}{(n_{11}+n_{10})n_{00}(n_{11}+n_{01})(n_{10}+n_{00})} = N\frac{n_{11}}{n_{11}+n_{10}}\frac{n_{11}}{n_{11}+n_{01}}\frac{n_{00}}{n_{10}+n_{00}} \tag{3.5}$$

으로 근사된다. $p_{01} \approx 0$이라는 가정 아래에서 로그함수의 테일러 전개를 이용하면

$$\begin{aligned}
p_{11}\log\frac{p_{11}}{p_{1\cdot}p_{\cdot 1}} &\approx p_{11}\log\frac{p_{11}}{(p_{11}+p_{10})p_{11}} \approx -p_{11}\log(1-p_{00}) \approx p_{11}\left(p_{00}+p_{00}^2\right),\\
p_{10}\log\frac{p_{10}}{p_{1\cdot}p_{\cdot 0}} &\approx p_{10}\log\frac{p_{10}}{(1-p_{00})(1-p_{11})} \approx p_{10}\log\left(1-\frac{p_{00}p_{11}}{(1-p_{00})(1-p_{11})}\right) \approx \frac{-p_{10}p_{00}p_{11}}{(1-p_{00})(1-p_{11})},\\
p_{01}\log\frac{p_{01}}{p_{0\cdot}p_{\cdot 1}} &\approx p_{01}\log\frac{p_{01}}{p_{00}p_{11}} \approx 0,\\
p_{00}\log\frac{p_{00}}{p_{0\cdot}p_{\cdot 0}} &\approx p_{00}\log\frac{p_{00}}{p_{00}(p_{00}+p_{10})} = -p_{00}\log(1-p_{11}) \approx p_{00}p_{11}
\end{aligned}$$

이다. 첫번째 근사식에서 p_{00}은 0에 가까운 값이 아니므로 $\log(1-p_{00})$을 이차 이상의 식으로 근사하였다. 이상과 같은 근사식을 이용하면 정보이득은 근사적으로

$$\mathrm{IG} \approx p_{00}p_{11}\left(2+p_{00}-\frac{p_{10}}{(1-p_{00})(1-p_{11})}\right) \tag{3.6}$$

로 표현된다.

식 (3.5)와 (3.6)에서 카이제곱통계량과 정보이득은 간단한 선형관계식으로 표현하기 어려움을 알 수 있다. 한편, Table 2에서 확인할 수 있는 것처럼 텍스트 데이터에서는 p_{00}이 p_{11}, p_{10}, p_{01}에 비해 매우 커서 1에 가까운 경우가 흔히 관찰되고 이런 경우에 카이제곱통계량은

$$\chi^2 \approx N\frac{p_{11}}{p_{11}+p_{10}}\frac{p_{11}}{p_{11}+p_{01}}\frac{p_{00}}{p_{10}+p_{00}} \approx N\frac{p_{11}}{p_{1\cdot}}\frac{p_{11}}{p_{\cdot 1}}$$

Table 2: Top-10 categories of the Reuters-21578 data

Category	earn	acq	money-fx	grain	crude	trade	interest	ship	wheat	corn
# of articles	3964	2369	717	582	578	485	478	286	283	237
($p_{1\cdot}$)	(0.44)	(0.26)	(0.08)	(0.06)	(0.06)	(0.05)	(0.05)	(0.03)	(0.03)	(0.03)
average p_{11}	0.010	0.008	0.004	0.003	0.003	0.003	0.002	0.001	0.001	0.001
average p_{10}	0.429	0.254	0.076	0.062	0.061	0.050	0.051	0.030	0.030	0.025
average p_{01}	0.022	0.023	0.027	0.029	0.028	0.028	0.029	0.030	0.030	0.030
average p_{00}	0.540	0.715	0.893	0.907	0.908	0.918	0.918	0.939	0.938	0.944

으로 주어지고 정보이득은

$$\mathrm{IG} \approx p_{00}p_{11}\left(2 + p_{00} - \frac{p_{10}}{(1-p_{00})(1-p_{11})}\right) \approx p_{11}\left(3 - \frac{p_{10}}{p_{1\cdot}}\right)$$

으로 근사되므로

$$\chi^2 \approx N\frac{p_{11}}{p_{\cdot 1}(3p_{1\cdot} - p_{10})}\,\mathrm{IG} \tag{3.7}$$

와 같이 간단한 근사관계식으로 표현할 수 있다.

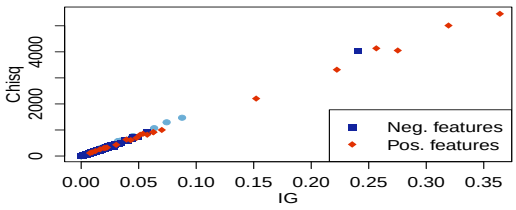
식 (3.7)에서 p_{11}, $p_{\cdot 1}$, $p_{1\cdot}$, p_{10}은 모두 0에 가까운 값들이므로 $p_{11}/p_{\cdot 1}(3p_{1\cdot} - p_{10})$은 1보다 큰 값이 된다. 따라서 식 (3.4)와 식 (3.7)을 비교해보면 정보이득 값이 동일하면 음변수일 때보다 양변수일 때 카이제곱통계량이 상대적으로 큰 값을 가진다는 것을 알 수 있다. 즉, 카이제곱통계량과 정보이득 값이 큰 단어부터 차례로 단어를 선택하는 경우 카이제곱통계량이 상대적으로 많은 양변수를 선택하고 정보이득은 카이제곱통계량에 비해 상대적으로 많은 음변수를 선택할 것이라는 것을 확인할 수 있다.

4. 실제 데이터에의 적용

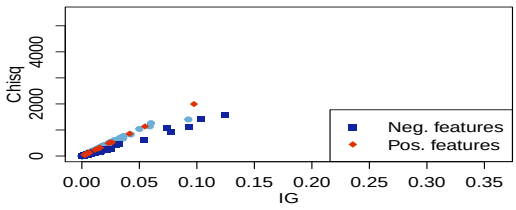
이 절에서는 지금까지 살펴본 카이제곱통계량과 정보이득의 특징을 실제 데이터를 통해 확인해보려 한다. 로이터-21578 데이터(Manning과 Schutze, 1999)는 로이터(Reuters) 통신사의 기사를 수집하여 정리한 데이터로 주로 무역, 기업, 금융시장 등과 관련된 주제들을 다룬 기사들로 구성되어 있다. 이 연구에서는 파이썬(Python) nltk (Natural Language Toolkit) 패키지(Bird 등, 2009)의 로이터-21578 데이터를 사용하였다.

이 데이터는 모두 10,788건의 기사로 이루어져 있고 90개의 범주로 기사들이 분류되어 있는데 많은 범주들이 소수의 기사로만 구성되어 있으므로 기사 건수 기준으로 상위 10개의 범주에 속하는 기사들에 대해서만 다루기로 한다. 해당 기사는 모두 9,034건으로 Table 2와 같다. 한 기사가 복수의 범주로 분류되어 있는 경우도 있어 표의 기사 건수 합계는 9,979이다. 표에 제시된 범주 중에서 “earn”은 기업의 이익을, “acq”는 기업 인수(acquisition)를, “money-fx”는 외환(foreign exchange)을 의미한다. 이 데이터에서 “earn”과 “acq” 이외의 주제들은 $p_{1\cdot}$ 값이 0.1 이하로 불균형 데이터에 해당되므로 앞서 살펴보았던 불균형 텍스트 데이터에서의 카이제곱통계량과 정보이득 사이의 관계가 성립할 것으로 예상할 수 있다.

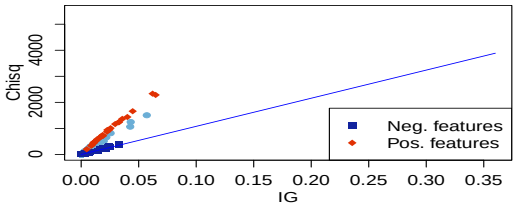
Figure 1은 Reuters-21578 데이터의 상위 10개 범주에 대한 카이제곱통계량과 정보이득의 산점도이다. 각 하위그림에서 붉은 색 마름모는 양변수를, 푸른 색 정사각형은 음변수를 나타내며 양변수는 오즈비가 20보다 큰 단어들로, 음변수는 오즈비가 0.2보다 작은 단어들로 선택하였다. Figure 1-(c)~(h)의 직선은 식 (3.4)에서 구한 카이제곱통계량과 정보이득 사이의 근사 선형관계식에 해당된다. Figure 1-(a)~(b)의 범주 “earn”과 “acq”의 경우 p_{10} 값이 커서 식 (3.4)와 같이 근사될 수 없으므로 직선을 생략하였다.



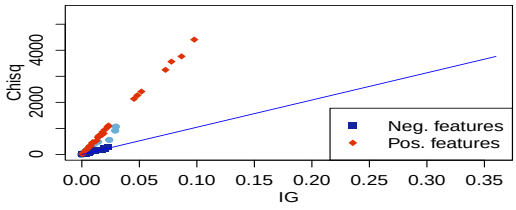
(a) earn



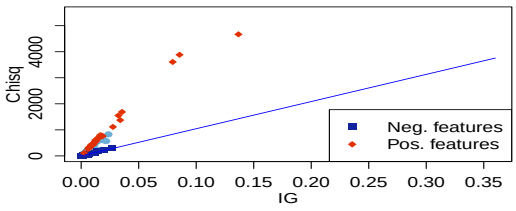
(b) acq



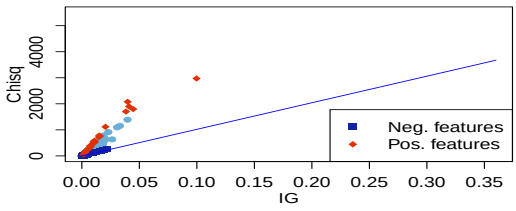
(c) money



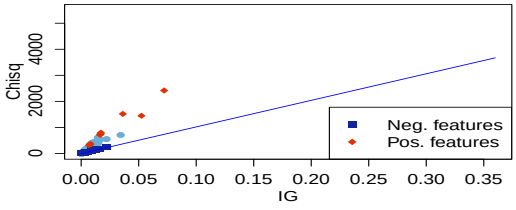
(d) grain



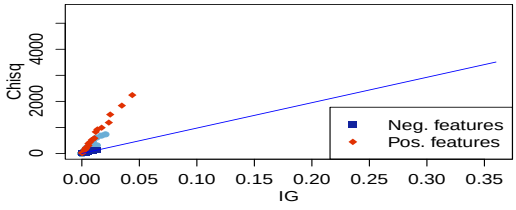
(e) crude



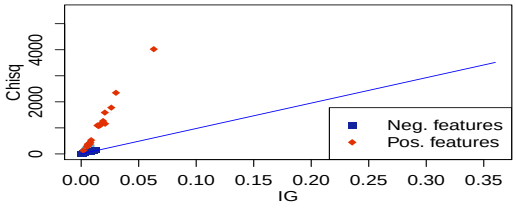
(f) trade



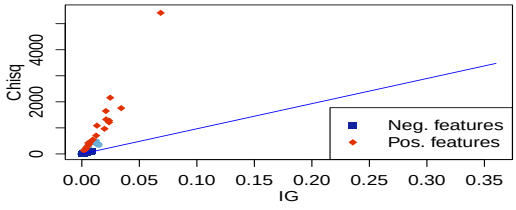
(g) interest



(h) wheat



(i) ship



(j) corn

Figure 1: *Comparison between chi-square statistic and information gain.*

Table 3: Number of positive features selected by chi-square statistic and information gain

		1 ~ 10	11 ~ 20	21 ~ 30	31 ~ 40	41 ~ 50	51 ~ 60	61 ~ 70	71 ~ 80	81 ~ 90	91 ~ 100
χ^2	earn	9	6	7	1	1	1	7	2	2	0
	acq	6	8	10	8	9	8	3	9	5	4
	money-fx	10	10	10	10	8	9	9	7	8	10
	grain	10	10	10	9	6	6	9	9	9	8
	crude	10	10	10	10	9	8	9	9	8	10
	trade	10	10	10	10	10	10	9	9	9	10
	interest	10	10	10	8	7	7	7	9	10	10
	wheat	10	10	10	10	6	5	10	8	10	9
	ship	10	10	10	8	9	7	8	9	9	8
	corn	10	10	10	10	6	9	7	9	8	9
IG	earn	9	7	4	3	0	2	2	4	2	3
	acq	5	8	8	7	6	7	7	4	5	5
	money-fx	9	6	8	8	9	10	9	9	9	5
	grain	10	5	8	7	9	9	2	5	5	7
	crude	9	9	7	9	10	8	10	6	10	8
	trade	10	7	8	9	10	8	9	10	10	10
	interest	8	8	6	7	9	7	7	6	7	10
	wheat	10	6	6	8	8	8	4	3	6	5
	ship	10	4	9	7	8	10	7	8	6	8
	corn	10	5	7	8	8	8	6	8	5	7

그림에서 살펴볼 수 있듯이 음변수의 경우 푸른 색 사각형으로 표현된 관측값들이 식 (3.4)를 이용하여 구한 직선에 가깝게 위치하고 있다. 즉, 음변수의 경우 식 (3.4)에서 살펴본 카이제곱통계량과 정보이득의 근사관계가 성립한다는 것을 확인할 수 있다.

한편, Figure 1-(a)~(b)의 경우 양변수와 음변수의 카이제곱통계량과 정보이득의 값 차이가 크지 않은 반면, 나머지 범주의 경우 양변수와 음변수의 차이가 상대적으로 큰 편이다. 즉, 범주 c에 속하는 문서가 많지 않은 경우에 비슷한 수준의 정보이득 값을 가지는 관측값들이 양변수인 경우 카이제곱통계량이 상대적으로 크게 나타나는 경향을 확인할 수 있다. 반대로 비슷한 수준의 카이제곱통계량 값을 가지는 단어들 중 양변수인 단어들의 정보이득 값이 상대적으로 작게 나타나는 경향이 있다. 따라서 정보이득을 사용하는 경우 카이제곱통계량을 사용하는 경우에 비해 음변수를 더 많이 선택하게 된다는 것을 알 수 있다.

Table 3에는 선택되는 단어 수를 10개 단위로 늘렸을 때 추가로 선택된 단어 중 양변수에 해당되는 단어의 수가 기록되어 있다. 표에서 선택된 단어 수가 30개일 때까지 카이제곱통계량은 두 개의 범주 “earn”, “acq”를 제외한 대부분의 범주에서 양변수만 선택하는 것을 알 수 있다. 반면, 정보이득은 카이제곱통계량에 비해 적은 수의 양변수, 즉 더 많은 수의 음변수를 선택하였다.

Table 4에는 카이제곱통계량과 정보이득에 의해 선택된 단어들 중 공통된 단어의 수가 기록되어 있다. 40개의 단어가 선택될 때까지는 카이제곱통계량으로 선택된 단어들과 정보이득으로 선택된 단어들에 일부 차이가 있지만 단어 수가 늘어날수록 차이가 작아져 100개의 단어를 선택했을 때는 “wheat”, “grain” 등 일부 범주를 제외한 대부분의 범주에서 90% 내외의 단어를 공유하는 것을 확인할 수 있다.

마지막으로 카이제곱통계량과 정보이득으로 선택된 단어들을 이용했을 때의 분류 성능을 확인해보았다. 로이터-21578 데이터의 문서 9,034건 중 3/4에 해당되는 문서들을 랜덤으로 추출하여 훈련용 데이터를 만들고 이 훈련용 데이터에 카이제곱통계량과 정보이득을 적용하여 단어들을 선택한 후 R 패키지 `rpart`를 이용하여 분류나무모형을 생성하였다. `minsplit=30`으로 지정하여 중간노드를 나누기 위해 필요한 최소한의 개체 수

Table 4: Number of common features (cumulative)

	1 ~ 10	1 ~ 20	1 ~ 30	1 ~ 40	1 ~ 50	1 ~ 60	1 ~ 70	1 ~ 80	1 ~ 90	1 ~ 100
earn	9	19	28	40	48	58	64	76	86	97
acq	9	19	27	36	43	52	66	71	81	92
money-fx	8	15	22	29	41	50	62	74	84	90
grain	8	14	23	31	44	57	60	65	72	81
crude	8	17	24	30	42	53	64	70	80	92
trade	9	15	24	31	41	49	54	66	77	86
interest	7	15	22	30	42	51	62	69	77	87
wheat	9	15	22	30	42	54	59	63	70	76
ship	9	13	22	30	39	53	61	69	78	87
corn	9	15	21	28	42	50	58	69	76	84

를 30으로 하고 `xval=10`으로 지정하여 10겹 교차검증을 하여 분류나무모형을 작성한 후 복잡도(complexity parameter) `cp=0.1` 수준에서 가지치기하여 최종 모형을 생성하였다. 이렇게 생성된 분류나무모형을 나머지 1/4에 해당되는 테스트 데이터에 적용하여 분류 성능을 평가하였다.

이상의 과정을 100번 반복하여 정확도(accuracy), 정밀도(precision), 민감도(sensitivity), 특이도(specificity)와 F1 지표를 구하여 분류 성능을 평가한 결과가 Tables 5와 6에 제시되어 있다. 표에서 살펴볼 수 있듯이 선택된 단어가 적을 때는 두 지표를 이용하여 선택한 단어들의 분류 성능에 다소 차이가 있다. 예를 들어 주제 "money-fx"에서 20개의 단어들이 선택되었을 때 카이제곱통계량으로 뽑힌 단어들을 이용한 분류 결과는 정밀도가 0.773, 민감도가 0.802인 반면, 정보이득으로 선택된 20개의 단어들을 이용한 분류모형은 정밀도 0.804, 민감도 0.735로 상당한 차이가 있다. 다만, 카이제곱통계량과 오즈비 중 어느 한 방법이 우월하다고 볼만한 뚜렷한 증거는 찾기 어렵다. 또 Table 4에서 확인한 바와 같이 선택되는 단어 개수가 많아질수록 두 지표에 의해 선택된 단어들 중 공통된 단어들이 점차 늘어나게되고 분류 성능의 차이도 축소되는 경향을 확인할 수 있다.

5. 결론 및 토의

이 연구에서는 텍스트 데이터에서 중요한 단어를 선택하기 위해 자주 사용되는 카이제곱통계량과 정보이득의 이론적 성질을 살펴보고 실제 데이터를 이용하여 이 성질들을 확인해보았다. 카이제곱통계량과 정보이득은 볼록성, 비음성 등의 성질을 공유하지만 불균형 데이터에서는 카이제곱통계량이 양변수 위주로 단어들을 선택하는 반면, 정보이득은 상대적으로 많은 음변수를 선택하는 차이점이 있음을 확인하였다. 이러한 결과는 Yang과 Pedersen (1997), Mladenić과 Grobelnik (1999), Forman (2003) 등의 선행연구에서 살펴볼 수 있는 두 지표의 성능 차이를 설명할 수 있는 근거가 될 수 있을 것으로 보인다. 로이터-21578 데이터에 두 지표를 적용하여 선택한 단어들을 이용하여 생성한 분류나무모형의 성능을 살펴본 결과 선택된 단어의 수가 적을 때 세부적으로 상당한 차이가 존재하는 경우가 있음을 확인하였다. 즉, 두 단어 선택지표 중 어떤 지표를 선택하는지에 따라 분류 결과에 실질적인 차이가 발생한다는 것을 확인하였다.

Son (2020)에서 제시된 바와 같이 오즈비는 양변수만 선택하는 단어 선택 지표이고 카이제곱통계량은 양변수와 음변수를 함께 선택할 수 있는 단어 선택 지표이지만 불균형 텍스트 데이터에서는 음변수보다 양변수를 많이 선택한다는 것이 알려져 있다. 이런 측면에서 정보이득이 세 지표 중 가장 많은 음변수를 선택하는 지표에 해당된다. 따라서 음변수를 함께 사용하는 것이 유용하다고 판단되는 경우에는 정보이득을, 양변수를 많이 사용하는 것이 유용하다고 판단되는 경우에는 카이제곱통계량이나 오즈비를 사용하는 것이 바람직하다고 볼 수 있다.

Table 5: Average classification performance : averaged for 100 random test data sets

Category			~ 10	~ 20	~ 30	~ 40	~ 50	~ 60	~ 70	~ 80	~ 90	~ 100
		accuracy	0.930	0.928	0.928	0.929	0.929	0.931	0.940	0.945	0.946	0.946
		precision	0.942	0.934	0.934	0.937	0.937	0.938	0.940	0.943	0.943	0.943
	χ^2	sensitivity	0.896	0.899	0.899	0.899	0.899	0.903	0.923	0.933	0.933	0.933
		specificity	0.956	0.950	0.950	0.953	0.953	0.953	0.954	0.955	0.956	0.956
earn		F1	0.918	0.916	0.916	0.917	0.917	0.920	0.932	0.938	0.938	0.938
		accuracy	0.917	0.928	0.928	0.929	0.929	0.929	0.932	0.939	0.944	0.946
		precision	0.919	0.934	0.934	0.937	0.937	0.937	0.938	0.939	0.942	0.943
	IG	sensitivity	0.890	0.899	0.899	0.899	0.899	0.899	0.904	0.920	0.931	0.933
		specificity	0.939	0.950	0.950	0.953	0.953	0.953	0.953	0.953	0.955	0.956
		F1	0.905	0.916	0.916	0.917	0.917	0.917	0.921	0.930	0.936	0.938
		accuracy	0.871	0.901	0.899	0.904	0.910	0.912	0.914	0.914	0.914	0.914
		precision	0.879	0.813	0.792	0.809	0.837	0.847	0.855	0.855	0.854	0.845
	χ^2	sensitivity	0.590	0.811	0.837	0.830	0.814	0.810	0.809	0.809	0.810	0.822
		specificity	0.971	0.933	0.921	0.930	0.944	0.948	0.951	0.951	0.951	0.947
acq		F1	0.705	0.811	0.813	0.818	0.825	0.828	0.831	0.831	0.831	0.833
		accuracy	0.868	0.901	0.904	0.909	0.913	0.914	0.914	0.914	0.915	0.914
		precision	0.839	0.812	0.808	0.835	0.852	0.855	0.855	0.855	0.849	0.842
	IG	sensitivity	0.623	0.812	0.831	0.815	0.809	0.809	0.809	0.810	0.819	0.828
		specificity	0.955	0.932	0.929	0.943	0.950	0.951	0.951	0.951	0.948	0.945
		F1	0.710	0.810	0.818	0.825	0.830	0.831	0.831	0.831	0.834	0.835
		accuracy	0.962	0.965	0.966	0.966	0.966	0.966	0.967	0.968	0.968	0.968
		precision	0.780	0.773	0.774	0.774	0.772	0.776	0.796	0.797	0.800	0.799
	χ^2	sensitivity	0.733	0.802	0.816	0.817	0.814	0.810	0.797	0.797	0.801	0.800
		specificity	0.982	0.980	0.979	0.979	0.979	0.980	0.982	0.982	0.983	0.982
money-fx		F1	0.755	0.786	0.794	0.794	0.792	0.792	0.796	0.796	0.799	0.799
		accuracy	0.964	0.965	0.967	0.968	0.968	0.968	0.968	0.968	0.968	0.968
		precision	0.787	0.804	0.796	0.801	0.801	0.799	0.800	0.798	0.798	0.796
	IG	sensitivity	0.749	0.735	0.783	0.801	0.801	0.803	0.802	0.804	0.805	0.806
		specificity	0.982	0.984	0.983	0.983	0.983	0.983	0.983	0.982	0.982	0.982
		F1	0.766	0.767	0.788	0.800	0.800	0.800	0.800	0.800	0.800	0.800
		accuracy	0.990	0.991	0.992	0.992	0.992	0.992	0.992	0.992	0.992	0.992
		precision	0.922	0.924	0.927	0.927	0.927	0.927	0.927	0.927	0.927	0.927
	χ^2	sensitivity	0.917	0.946	0.944	0.944	0.944	0.944	0.944	0.944	0.944	0.944
		specificity	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995
grain		F1	0.919	0.935	0.936	0.936	0.936	0.936	0.936	0.936	0.936	0.936
		accuracy	0.989	0.991	0.992	0.992	0.992	0.992	0.992	0.992	0.992	0.992
		precision	0.920	0.927	0.927	0.927	0.927	0.927	0.927	0.927	0.927	0.927
	IG	sensitivity	0.914	0.933	0.944	0.944	0.944	0.944	0.944	0.944	0.944	0.944
		specificity	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995
		F1	0.917	0.930	0.936	0.936	0.936	0.936	0.936	0.936	0.936	0.936
		accuracy	0.977	0.977	0.977	0.977	0.977	0.977	0.977	0.977	0.977	0.977
		precision	0.820	0.834	0.834	0.836	0.837	0.839	0.839	0.839	0.840	0.839
	χ^2	sensitivity	0.837	0.818	0.818	0.816	0.816	0.814	0.815	0.815	0.814	0.814
		specificity	0.987	0.988	0.988	0.988	0.989	0.989	0.989	0.989	0.989	0.989
crude		F1	0.825	0.822	0.822	0.821	0.823	0.823	0.823	0.823	0.823	0.823
		accuracy	0.977	0.977	0.977	0.977	0.977	0.977	0.977	0.977	0.977	0.978
		precision	0.804	0.831	0.837	0.840	0.839	0.838	0.837	0.836	0.835	0.835
	IG	sensitivity	0.854	0.820	0.815	0.814	0.815	0.816	0.816	0.816	0.817	0.821
		specificity	0.985	0.988	0.989	0.989	0.989	0.989	0.989	0.989	0.988	0.988
		F1	0.827	0.822	0.822	0.823	0.823	0.823	0.823	0.823	0.823	0.825

한편, 향후 적절한 단어의 개수를 찾기 위한 통계적 방법에 대한 연구가 필요한 것으로 보인다. Tables 5와 6에서 일부 범주에서는 단어의 개수를 늘림에 따라 분류 성능이 개선되지 않거나 오히려 나빠지는 현상을

Table 6: Average classification performance : averaged for 100 random test data set

Category			~ 10	~ 20	~ 30	~ 40	~ 50	~ 60	~ 70	~ 80	~ 90	~ 100
		accuracy	0.975	0.975	0.976	0.975	0.975	0.975	0.975	0.975	0.975	0.975
		precision	0.799	0.802	0.803	0.801	0.799	0.797	0.796	0.792	0.794	0.796
	χ^2	sensitivity	0.716	0.720	0.721	0.721	0.722	0.720	0.720	0.719	0.717	0.712
		specificity	0.990	0.990	0.990	0.990	0.990	0.990	0.990	0.989	0.989	0.990
trade		F1	0.754	0.757	0.758	0.758	0.757	0.756	0.755	0.753	0.752	0.751
		accuracy	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975
		precision	0.799	0.789	0.794	0.798	0.798	0.798	0.797	0.798	0.801	0.801
	IG	sensitivity	0.716	0.718	0.720	0.712	0.712	0.712	0.712	0.710	0.709	0.708
		specificity	0.990	0.989	0.989	0.990	0.990	0.990	0.990	0.990	0.990	0.990
		F1	0.754	0.750	0.754	0.751	0.751	0.751	0.751	0.750	0.751	0.750
		accuracy	0.968	0.968	0.969	0.969	0.969	0.970	0.974	0.974	0.974	0.974
		precision	0.770	0.732	0.730	0.731	0.726	0.736	0.782	0.785	0.785	0.785
	χ^2	sensitivity	0.573	0.642	0.665	0.665	0.665	0.684	0.704	0.703	0.701	0.701
		specificity	0.990	0.987	0.986	0.986	0.986	0.986	0.989	0.989	0.989	0.989
interest		F1	0.655	0.681	0.693	0.694	0.692	0.707	0.740	0.740	0.739	0.739
		accuracy	0.964	0.968	0.968	0.972	0.974	0.974	0.974	0.974	0.974	0.974
		precision	0.734	0.725	0.711	0.762	0.784	0.785	0.785	0.785	0.784	0.785
	IG	sensitivity	0.519	0.651	0.681	0.700	0.704	0.703	0.703	0.702	0.702	0.700
		specificity	0.989	0.986	0.984	0.988	0.989	0.989	0.989	0.989	0.989	0.989
		F1	0.603	0.683	0.693	0.728	0.740	0.740	0.740	0.740	0.739	0.739
		accuracy	0.994	0.994	0.994	0.994	0.994	0.994	0.994	0.994	0.994	0.994
		precision	0.861	0.861	0.861	0.861	0.861	0.861	0.861	0.861	0.863	0.864
	χ^2	sensitivity	0.965	0.965	0.965	0.965	0.965	0.965	0.965	0.964	0.961	0.957
		specificity	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995
wheat		F1	0.910	0.910	0.910	0.910	0.910	0.910	0.910	0.909	0.909	0.908
		accuracy	0.994	0.994	0.994	0.994	0.994	0.994	0.994	0.994	0.994	0.994
		precision	0.861	0.861	0.861	0.861	0.861	0.862	0.862	0.864	0.865	0.866
	IG	sensitivity	0.965	0.965	0.965	0.965	0.965	0.964	0.962	0.959	0.956	0.955
		specificity	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995
		F1	0.910	0.910	0.910	0.910	0.910	0.909	0.909	0.908	0.908	0.908
		accuracy	0.986	0.986	0.986	0.986	0.986	0.986	0.986	0.986	0.986	0.986
		precision	0.854	0.855	0.855	0.855	0.831	0.807	0.807	0.807	0.807	0.807
	χ^2	sensitivity	0.679	0.680	0.679	0.678	0.711	0.746	0.742	0.740	0.739	0.739
		specificity	0.996	0.996	0.996	0.996	0.995	0.994	0.994	0.994	0.994	0.994
ship		F1	0.755	0.755	0.755	0.755	0.763	0.773	0.771	0.771	0.770	0.770
		accuracy	0.986	0.986	0.986	0.986	0.986	0.986	0.986	0.986	0.986	0.986
		precision	0.854	0.847	0.803	0.796	0.799	0.804	0.805	0.806	0.807	0.806
	IG	sensitivity	0.679	0.682	0.746	0.753	0.750	0.746	0.742	0.740	0.739	0.738
		specificity	0.996	0.996	0.994	0.994	0.994	0.994	0.994	0.994	0.994	0.994
		F1	0.755	0.753	0.772	0.772	0.772	0.772	0.770	0.770	0.770	0.769
		accuracy	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995
		precision	0.870	0.869	0.869	0.869	0.868	0.868	0.867	0.869	0.870	0.870
	χ^2	sensitivity	0.959	0.947	0.947	0.947	0.950	0.951	0.953	0.955	0.961	0.961
		specificity	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996
corn		F1	0.911	0.905	0.905	0.905	0.905	0.906	0.907	0.909	0.912	0.912
		accuracy	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.995
		precision	0.869	0.868	0.868	0.867	0.869	0.871	0.871	0.871	0.871	0.871
	IG	sensitivity	0.947	0.950	0.952	0.954	0.958	0.960	0.960	0.961	0.961	0.961
		specificity	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996
		F1	0.904	0.905	0.906	0.907	0.910	0.912	0.912	0.913	0.913	0.913

확인할 수 있으며 범주마다 적절한 단어의 개수에 차이가 있다는 것도 알 수 있다. 카이제곱통계량과 정보이득의 경우 동질성에 대한 귀무가설 아래에서 점근적으로 카이제곱분포를 따른다는 사실이 알려져 있지만

이러한 가설검정 절차를 이용하여 단어 수를 선택하면 지나치게 많은 단어들이 유의미한 단어로 선택되는 경우를 볼 수 있다. 따라서 실제 데이터 분석에서는 통계적인 가설검정 절차를 이용하는 대신 주관적인 기준으로 단어 수를 결정하는 것이 일반적이다. 이렇게 통계적인 가설검정 절차가 잘 작동하지 않는 것은 Rao와 Scott (1987)이 제시한 바와 같이 편향(bias)에 기인한 것으로 볼 수 있으므로 Rao-Scott 카이제곱검정 절차 등 편향을 보정한 가설검정 절차도 고려해 볼 필요가 있다.

Appendix: 카이제곱통계량과 정보이득의 볼록성에 대한 증명

식 (2.2)에서 카이제곱통계량은

$$\chi^2 = \sum_{i=0}^{1}\sum_{j=0}^{1}\frac{(n_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i=0}^{1}\sum_{j=0}^{1}\frac{(Np_{ij}-Np_{i\cdot}p_{\cdot j})^2}{Np_{i\cdot}p_{\cdot j}} = \sum_{i=0}^{1}\sum_{j=0}^{1}\frac{N(p_{ij}-p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}}$$

과 같이 나타낼 수 있다. 여기서 N은 상수이므로 볼록함수임을 보이기 위해서는 고려하지 않아도 된다. 먼저 카이제곱통계량의 두 항

$$\frac{(p_{11}-p_{1\cdot}p_{\cdot 1})^2}{p_{1\cdot}p_{\cdot 1}} = \frac{\{p_{11}-p_{1\cdot}(p_{11}+p_{01})\}^2}{p_{1\cdot}(p_{11}+p_{01})},$$
$$\frac{(p_{10}-p_{1\cdot}p_{\cdot 0})^2}{p_{1\cdot}p_{\cdot 0}} = \frac{\{p_{1\cdot}-p_{11}-p_{1\cdot}(p_{1\cdot}-p_{11}+p_{00})\}^2}{p_{1\cdot}(p_{1\cdot}-p_{11}+p_{00})}$$

은 $x = p_{11}$과 상수 a, b, c에 대해 함수

$$f(x) = \frac{\{-x+a+c(x+a+b)\}^2}{c(x+a+b)}$$

형태로 표현할 수 있으며 $U = u+a+b$, $V = v+a+b$, $d = 2a+b$라 하면

$$f(u) = \frac{\{-u+a+c(u+a+b)\}^2}{c(u+a+b)} = \frac{(-U+d+cU)^2}{cU} = \frac{\{(c-1)U+d\}^2}{cU},$$
$$f(v) = \frac{\{-v+a+c(v+a+b)\}^2}{c(v+a+b)} = \frac{(-V+d+cV)^2}{cV} = \frac{\{(c-1)V+d\}^2}{cV}$$

이고

$$\begin{aligned} f(\lambda u+(1-\lambda)v) &= \frac{\{-(\lambda u+(1-\lambda)v)+a+c((\lambda u+(1-\lambda)v)+a+b)\}^2}{c((\lambda u+(1-\lambda)v)+a+b)} \\ &= \frac{\{-\lambda U-(1-\lambda)V+d+c((\lambda U+(1-\lambda)V)\}^2}{c(\lambda U+(1-\lambda)V)} \\ &= \frac{\{\lambda((c-1)U+d)+(1-\lambda)((c-1)V+d)\}^2}{\lambda cU+(1-\lambda)cV} \end{aligned}$$

이므로

$$f(\lambda u+(1-\lambda)v) \le \lambda f(u)+(1-\lambda)f(v)$$

임을 알 수 있다.

마찬가지로 카이제곱통계량의 두 항

$$\frac{(p_{01} - p_{0\cdot}p_{\cdot 1})^2}{p_{0\cdot}p_{\cdot 1}} = \frac{\{p_{01} - p_{0\cdot}(p_{11} + p_{01})\}^2}{p_{0\cdot}(p_{11} + p_{01})},$$
$$\frac{(p_{00} - p_{0\cdot}p_{\cdot 0})^2}{p_{0\cdot}p_{\cdot 0}} = \frac{\{p_{0\cdot} - p_{0\cdot}(p_{1\cdot} - p_{11} + p_{00})\}^2}{p_{0\cdot}(p_{1\cdot} - p_{11} + p_{00})}$$

은 $x = p_{11}$과 상수 a, b, c에 대해 함수

$$g(x) = \frac{(ax + b + c)^2}{ax + b}$$

형태로 표현할 수 있으며 $U = au + b$, $V = av + b$라 하면

$$g(u) = \frac{(au + b + c)^2}{au + b} = \frac{(U + c)^2}{U}, \qquad g(v) = \frac{(av + b + c)^2}{av + b} = \frac{(V + c)^2}{V}$$

이고

$$g(\lambda u + (1 - \lambda)v) = \frac{\{a(\lambda u + (1 - \lambda)v) + b + c\}^2}{a(\lambda u + (1 - \lambda)v) + b} = \frac{\{\lambda U + (1 - \lambda)V + c\}^2}{\lambda U + (1 - \lambda)V}$$

이므로

$$g(\lambda u + (1 - \lambda)v) \le \lambda g(u) + (1 - \lambda)g(v)$$

임을 알 수 있다. 따라서 카이제곱통계량의 각 항들이 p_{11}의 볼록함수이므로 카이제곱통계량도 p_{11}의 볼록함수이다.

정보이득의 볼록성은 로그합 부등식(log sum inequality)을 이용하여 보일 수 있다. 정보이득은

$$\mathrm{IG}(c, w) = \sum_{i=0,1}\sum_{j=0,1} p_{ij} \log \frac{p_{ij}}{p_{i\cdot}p_{\cdot j}} = p_{11} \log \frac{p_{11}}{p_{1\cdot}p_{\cdot 1}} + p_{10} \log \frac{p_{10}}{p_{1\cdot}p_{\cdot 0}} + p_{01} \log \frac{p_{01}}{p_{0\cdot}p_{\cdot 1}} + p_{00} \log \frac{p_{00}}{p_{0\cdot}p_{\cdot 0}}$$

과 같이 표현될 수 있다. 이 식에서 세 번째 항과 네 번째 항은

$$p_{01} \log \frac{p_{01}}{p_{0\cdot}p_{\cdot 1}} = -\log(p_{11} + p_{01}) + C_3,$$
$$p_{00} \log \frac{p_{00}}{p_{0\cdot}p_{\cdot 0}} = -\log(p_{1\cdot} - p_{11} + p_{00}) + C_4$$

이므로 $0 < p_{11} < p_{1\cdot}$에서 p_{11}의 볼록함수이다. 다음으로 첫번째 항은

$$p_{11} \log \frac{p_{11}}{p_{1\cdot}p_{\cdot 1}} = p_{11} \log \frac{p_{11}}{p_{1\cdot}(p_{11} + p_{01})}$$

으로 나타낼 수 있다. 로그합 부등식(log sum inequality)에서 양의 실수 a_i와 b_i $(i = 1, \ldots, k)$에 대해

$$\sum_i \left(a_i \log \frac{a_i}{b_i}\right) \ge \left(\sum_i a_i\right) \log \frac{\sum_i a_i}{\sum_i b_i}$$

가 성립하므로

$$\lambda u \log \frac{u}{p_{1\cdot}(u + p_{01})} + (1 - \lambda)v \log \frac{v}{p_{1\cdot}(v + p_{01})} \ge \{\lambda u + (1 - \lambda)v\} \log \frac{\lambda u + (1 - \lambda)v}{p_{1\cdot}\{\lambda u + (1 - \lambda)v + p_{01}\}}$$

이고 p_{11}의 볼록함수이다. 마찬가지로 두번째 항은

$$p_{10}\log\frac{p_{10}}{p_{1\cdot}p_{\cdot 0}} = p_{10}\log\frac{p_{10}}{p_{1\cdot}(p_{10}+p_{00})} = (p_{1\cdot}-p_{11})\log\frac{(p_{1\cdot}-p_{11})}{p_{1\cdot}\{(p_{1\cdot}-p_{11})+p_{00}\}}$$

이고

$$\begin{aligned}
&\lambda(p_{1\cdot}-u)\log\frac{p_{1\cdot}-u}{p_{1\cdot}\{(p_{1\cdot}-u)+p_{01}\}} + (1-\lambda)(p_{1\cdot}-v)\log\frac{p_{1\cdot}-v}{p_{1\cdot}\{(p_{1\cdot}-v)+p_{01}\}} \\
&\geq \{\lambda(p_{1\cdot}-u)+(1-\lambda)(p_{1\cdot}-v)\}\log\frac{\lambda(p_{1\cdot}-u)+(1-\lambda)(p_{1\cdot}-v)}{p_{1\cdot}\{\lambda(p_{1\cdot}-u)+(1-\lambda)(p_{1\cdot}-v)+p_{01}\}} \\
&= \{p_{1\cdot}-(\lambda u+(1-\lambda)v)\}\log\frac{p_{1\cdot}-(\lambda u+(1-\lambda)v)}{p_{1\cdot}\{p_{1\cdot}-(\lambda u+(1-\lambda)v)+p_{01}\}}
\end{aligned}$$

이므로 p_{11}의 볼록함수이다. 이렇게 각 항들이 p_{11}의 볼록함수이므로 정보이득도 p_{11}의 볼록함수이다.

References

Bird S, Klein E, and Loper E (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly.

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

Kou G, Yang P, Peng Y, Xiao F, Chen Y, and Alsaadi FE (2020). Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Applied Soft Computing*, **86**, 105836.

Manning C and Schutze H (1999). *Foundations of statistical natural language processing*. MIT press.

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

Rao JNK and Scott AJ (1987). On simple adjustments to chi-square tests with sample survey data. *The annals of statistics*, **15**(1), 385–397.

Son W (2020). Skewness of chi-square statistic for imbalanced text data. *Journal of the Korean Data & Information Science Society*, **31**(5), 807–821.

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

Received March 2, 2022; Revised April 19, 2022; Accepted May 23, 2022

불균형 텍스트 데이터의 변수 선택에 있어서의 카이제곱통계량과 정보이득의 특징

문혜인[a], 손원[1,a]

[a]단국대학교 대학원 응용통계학과

요 약

텍스트 데이터는 일반적으로 많은 단어로 이루어져 있으므로 변수의 수가 매우 많은 고차원 데이터에 해당된다. 이러한 고차원 데이터에서는 계산 효율성과 통계분석의 정확성을 높이기 위해 많은 변수 중 중요한 변수를 선택하기 위한 절차를 거치는 경우가 많다. 텍스트 데이터에서도 많은 단어 중 중요한 단어를 선택하기 위해 여러가지 방법들이 사용되고 있다. 이 연구에서는 단어 선택을 위한 대표적인 필터링 방법인 카이제곱통계량과 정보이득의 공통점과 차이점을 살펴보고 실제 텍스트 데이터에서 이 단어선택 방법들의 성질을 확인해보았다. 카이제곱통계량과 정보이득은 비음성, 볼록성 등의 성질을 공유하지만 불균형 텍스트 데이터에서 카이제곱통계량이 양변수 위주로 단어를 선택하는 반면, 정보이득은 음변수도 상대적으로 많이 선택하는 경향이 있음을 확인하였다.

주요용어: 변수 선택, 불균형 데이터, 정보이득, 카이제곱통계량, 텍스트 데이터

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