문서 분류에서 단어의 통계 정보를 이용한 특징 선택 기법의 비교

-Comparison of Feature Selection Methods using the Statistics of Words in Text Categorization-

임윤택
Rim Yun-Taek
윤충화
Yoon Chung-Hwa

요 약

정보 검색 분야의 문서 분류에 기계 학습 기법을 적용할 때 발생하는 가장 큰 문제는 문서를 패턴으로 표현할 때, 하나의 패턴이 가지는 특징의 수가 기계 학습 기법에서 처리할 수 있는 범위를 넘어서는 것이다. 이러한 문제를 해결하기 위하여 특징 선택 기법은 패턴을 구성하고 있는 특징 중에서 실제 문서 분류에 많은 영향을 주는 특징만을 선택하여, 기계 학습 기법에서 쉽게 처리할 수 있을 정도의 패턴을 구성하게 한다.

본 논문에서는 이러한 특징 선택 기법 중에서 IG(Information Gain), Gini index, Relief-F, DF(Document Frequency)를 비교하였다. 실험 결과 문서들에 포함된 모든 고유 단어를 특징의 길이로 하여 패턴을 구성했을 때보다 특징 선택 기법을 적용하여 고유 단어 중 일부를 특징으로 패턴을 구성할 때 기계학습에서 더 향상된 분류 성능을 보였다.

1. 서론

정보 검색 분야에서 문서 분류는 문서를 미리 정의해 놓은 클래스로 분류하는 것을 말하며, 현재 기계학습 기법들을 정보 검색의 문서 분류 영역에 적용하려는 많은 연구가 진행 중이다.

기계 학습 기법을 문서 분류에 적용하기 위하여 문서를 패턴의 형태로 표현하게 되는데, 이 때, 문서들에 포함된 모든 고유 단어를 특징으로 하여 패턴이 구성 되며, 특징의 개수는 고유 단어의 개수와 일치하게 된다. 일반적으로 어느 정도의 내용을 가지는 텍스트 문서들에 포함된 고유 단어는 수 천에서 수 만개에 이르게 되며 이러한 문서들을 패턴으로 구성할 경우 기계 학습 기법은 수 천에서 수 만개의 특징을 가지는 패턴

명지대학교 컴퓨터공학과

을 처리해야 한다. 그러나, 대부분의 기계 학습 알고리즘의 경우 이와 같이 큰 수의 특징을 가지는 패턴을 처리하는 것은 어려운 문제이다.

이러한 문제를 해결하기 위해서 패턴을 구성하는 특징의 개수를 줄이는 특징 추출(Feature Extraction) 기법과 특징 선택(Feature Selection) 기법을 적용 할 수 있다. 특징 추출 기법은 몇 개의 특징을 조합하여 새로운 특징을 구성하는 것을 말하며, 특징 선택은 기법은 고유 단어의 통계를 고려하여 문서 분류에 작은 영향을 미치는 단어를 특징에서 제거하는 방법을 말한다.

본 논문에서는 기계학습 기법을 이용하여 효율적인 문서 분류가 가능하도록 하기 위해서 특징 선택 기법인 IG(Information Gain), Gini index, Relief-F, DF(Document Frequency)를 비교 분석하였다.

2. 특징 선택 기법

앞에서 언급한 바와 같이 특징 선택을 위해서 사용될 수 있는 기법으로는 IG, Gini index, Relief-F, DF등이 있다. 각 기법들은 모두 문서 안에서 고유 단어의 출현 횟수, 클래스의 출현 확률 등의 통계적 자료를 이용하여 각 고유 단어들에 대한 수치를 계산한다.

2.1 Information Gain(IG)

IG[1,2]는 기계학습 분야에서 학습 패턴에 소속된 특징의 변별력을 구하기 위해 자주 사용되는 기법이며, 한 문서 안에 어떤 단어가 존재하느냐 존재하지 않느냐를 사용하여 클래스를 추론하기위한 정보의 양을 계산한다. 분류해야 할 클래스가 n개일 때, 고유 단어 t의 IG값은 다음 (식1)을 이용하여 구한다.

$$\begin{aligned} G(t) = & -\sum_{i=1}^{n} P(c_i)\log P(c_i) \\ & + P(t)\sum_{i=1}^{n} P(c_i \mid t)\log P(c_i \mid t) \\ & + P(\bar{t})\sum_{i=1}^{n} P(c_i \mid \bar{t})\log P(c_i \mid \bar{t}) \end{aligned} \quad (1)$$

이 때, $P(ci)$는 전체 문서에서 클래스 ci가 발생할 확률을 말하고, $P(t)$는 전체 문서에서 단어 t가 발생할 확률을 말한다. 그리고, $P(ci|t)$는 단어 t가 발생했을 때, 클래스 ci에 소속될 확률을 말한다. $P(ci|t)$는 표1과 같은 단어 t의 two-way contingency table를 이용하여 구할 수 있다. 표 1에서 현재 IG를 계산하고 있는 단어 t는 $w+x$로 나타내고, 그 중 클래스 ci에 소속된 단어는 w로 나타낼 수 있다. 그러므로 $P(ci|t)$는 $w/w+x$이다.

표 1 Two-way continqency table

	C_i	$\tilde{}C_i$
t	w	x
$\tilde{}t$	y	z

주어진 문서에서 각각의 고유 단어들에 대한 IG값을 구하고, IG값에 의해 내림차순으로 정렬한 후에, 사전에 정의한 비율 내에 속하는 단어들 만을 선택하여 패턴을 구성하는 특징으로 사용한다.

2.2 Gini

Gini[3]는 IG와 유사한 기법으로 클래스를 추론하기 위한 정보의 양을 계산한다.

$$\begin{aligned} Gini(t) = & -\sum_{i=1}^{n} P(c_i)^2 \\ & + P(t)\sum_{i=1}^{n} P(c_i \mid t)^2 \\ & + P(\bar{t})\sum_{i=1}^{n} P(c_i \mid t)^2 \end{aligned} \tag{2}$$

IG와 마찬가지로, 각각 고유 단어들에 대한 Gini값을 구한 후에 사전에 정의한 비율 내에 속하는 단어들 만을 선택하여 패턴을 구성하는 특징으로 사용한다.

2.3 Relief-F

Relief[3]는 기계학습 분야에서 학습 패턴에 소속된 특징의 변별력을 구하기 위해 자주 사용되는 기법이다. 원래 Relief는 클래스가 두 개로 분할되는 문제에서만 사용할 수 있기 때문에 본 논문에서는 두 개 이상의 클래스로 분할하는 문제를 처리할 수 있는 Relief-F[3]를 사용하였다.

Relief-F의 주요 아이디어는 학습 패턴에 소속된 특징 중 주어진 특징의 한 값이 자신이 소속된 클래스를 얼마나 잘 구분해 낼 수 있는가를 추정하는 것이다. 이러한 목적을 달성하기 위해서 (식3)을 사용하여 각 고유 단어들에 대한 Relief-F의 값을 계산한다.

$$W[t] = \frac{\left(P(t)^2 + P(\bar{t})^2\right) \times Gini'(t)}{\sum_{i=1}^{n} P(c_i)^2 \times \left(1 - \sum_{i=1}^{n} P(c_i)^2\right)} \tag{3}$$

이 때, Gini(f)는 다음 (식4)와 같이 계산되고, 이 식이 Gini를 계산하는 (식 2)와 유사하기 때문에 Relief-F는 IG와 Gini의 식과 밀접한 관계가 있다.

$$Gini'(t) = -\sum_{i=1}^{n} P(c_i)^2 + \frac{P(t)^2}{P(t)^2 + P(\bar{t})^2} \sum_{i=1}^{n} P(c_i \mid t)^2 + \frac{P(\bar{t})^2}{P(t)^2 + P(\bar{t})^2} \sum_{i=1}^{n} P(c_i \mid t)^2 \quad (4)$$

IG와 마찬가지로, 각각 고유 단어들에 대한 Relief-F값을 구한 후에 사전에 정의한 비율 내에 속하는 단어들 만을 선택하여 패턴을 구성하는 특징으로 사용한다.

2.4 Document Frequency(DF)

DF[4]는 특징의 차원을 축소하기 위해서 사용되는 가장 간단한 기법 중 하나이다. DF는 한 단어가 전체 문서에서 출현한 횟수를 말한다. 학습으로 사용되는 문서에서 각각의 고유 단어들에 대한 DF를 계산한 후, 각각의 고유 단어들을 DF값에 의해 내림차순으로 정렬하여, 사전에 정의한 고유 단어의 비율 내에 속하는 상위 m%을 제외한 나머지 단어들을 특징 공간에서 제거한다.

3. 학습 기법

본 논문에서 4가지 특징 선택 기법으로 만들어진 패턴의 분류 성능을 비교하기 위해서 NN(Nearest Neighbors)[5] 분류기를 사용하였다. NN분류기는 메모리기반 학습기법을 사용한 최초의 분류기로 이 방법은 Lazy Learning Algorithm이라고도 하는데, 그 이유는 학습 시에는 단순히 학습 패턴을 메모리에 저장하며, 차후 입력패턴의 분류 시 모든 계산이 수행되기 때문이다

이러한 NN분류기의 개략적인 알고리즘은 다음과 같다.

① 주어진 학습패턴을 모두 메모리에 저장한다.
② 입력패턴 Q의 분류를 위하여 메모리에 저장된 모든 학습패턴과의 거리를 다음 (식 5)를 이용하여 계산한다.

$$D_{EQ} = \sqrt{\sum_{i=1}^{n} (E_{f_i} - Q_{f_i})^2} \quad (5)$$

이때 E는 메모리에 저장된 학습패턴을 나타내며, Q는 주어진 입력 패턴이다. 또한 n은 패턴을 구성하는 특징의 개수이며, 는 각각 학습패턴과 입력패턴의 i 번째 특징값을 나타낸다.
③ 입력패턴 Q와 가장 가까운 학습패턴을 선정한다.
④ 선택된 학습패턴이 소속되는 클래스로 입력패턴 Q를 분류한다.

4. 실험 결과 및 분석

본 논문에서는 4가지 특징 선택 기법을 로이터-21578[1]데이터를 이용하여 비교하였다. 전체 고유 단어를 특징의 개수로 하여 구성된 패턴과, 특징 선택 기법을 사용하여 선택된 특징으로 구성된 패턴의 분류 성능을 비교하기 위해서 NN분류기를 사용하였고, 학습 데이터와 테스트 데이터의 분할은 로이터-21578에서 제공하고 있는 수정된 Apte[1]분할을 따랐다. 수정된 Apte분할은 이전에 로이터-21578문서에서 제공하고 있는 3가지 분할 형태 중 하나로, 이전의 연구자들과의 연구 결과를 비교하기 위해서 다른 형태의 분할을 사용하지 않고, 제공된 분할 형태 중 하나를 선택하여 사용하였다.

로이터 문서를 기계학습에서 사용할 수 있는 형태의 패턴으로 바꾸기 위해서 stop-wording과 stemming의 전처리 과정을 수행하였다. stop-wording은 문서 내의 단어에서 불용어를 제거하는 과정이고, stemming은 형태적으로 유사한 단어를 모아 하나의 단어로 만드는 과정이다[7]. 다음으로, 텍스트 문서를 역 파일의 형태로 바꾼다. 역 파일은 모든 텍스트 문서에 포함된 고유 단어를 인덱스로 하여 원본 문서를 검색할 수 있는 구조로 되어 있으며, 문서 안에서 단어의 출현 횟수, 클래스별 단어의 수 등의 각각 고유 단어들에 대한 통계적 정보도 포함 된다. 이러한 방법으로 특징 선택 이전의 모든 특징들을 사용하는 학습 패턴을 구성할 경우 각 패턴은 19588개의 특징을 가지게 된다.

실험은 Windows NT를 적재한 PentiumII-333 컴퓨터를 사용하였으며, 모든 실험결과는 5회 반복측정 한 후 평균값으로 나타내었다.

4.1 실험 방법

실험 방법은 우선 특징 선택 기법이 NN분류기를 이용하여 클래스를 분류 하였을 때, 분류 성능에 어느 정도 영향을 주는가를 알아보기 위해서 문서의 모든 고유 단어를 특징의 개수로 하는 패턴과, IG값을 이용하여 고유 단어 중 일부를 선택한 후, 선택된 단어를 특징의 개수로 하는 패턴의 분류 성능을 비교 하였다. 이 때, IG값이 높은 단어 순으로 전체 고유 단어의 0.1%서 50%까지 늘려가면서 특징을 선택한 후, 각각 선택된 특징의 개수가 다를 때 분류 성능에 어떠한 영향을 주는가를 알아보았다. 각 특징은 TfIdf기법으로 표현하였다.

다음으로 앞의 실험에서 가장 좋은 분류 성능을 보인 1%의 단어를 특징으로 선택하는 것과, 선택된 특징을 TfIdf기법으로 표현하는 방법을 사용하여 IG, Gini, Relief-F, DF의 네 가지 특징 선택 기법으로 선택된 단어를 특징의 개수로 하는 패턴의 분류 성능을 비교하였다.

4.2 분류 성능 실험

그림 1의 분류 성능을 보면 전체 고유 단어를 특징의 개수로 패턴을 구성할 때와, IG기법을 사용하여 전체 단어 중 50%를 선택하여 패턴을 구성할 때, 분류 성능에 큰

차이를 보이지 않는다. 그러나, 전체 고유 단어를 특징의 개수로 할 때 패턴의 길이가 고유 단어의 수인 19,588인데 반하여 50%를 선택하여 구성된 패턴의 길이는 이것의 절반인 9,794이므로, 분류에 소요되는 시간에 현격한 차이를 보이게 된다. 또한 전체 단어 중 1%인 195개의 단어를 선택하였을 때, 가장 높은 분류 성능을 나타내는 것으로 미루어, 특징 선택 기법을 사용하여 분류에 영향을 많이 주는 단어만을 선택하였을 때 오히려 분류 성능이 향상된다는 것을 알 수 있다.

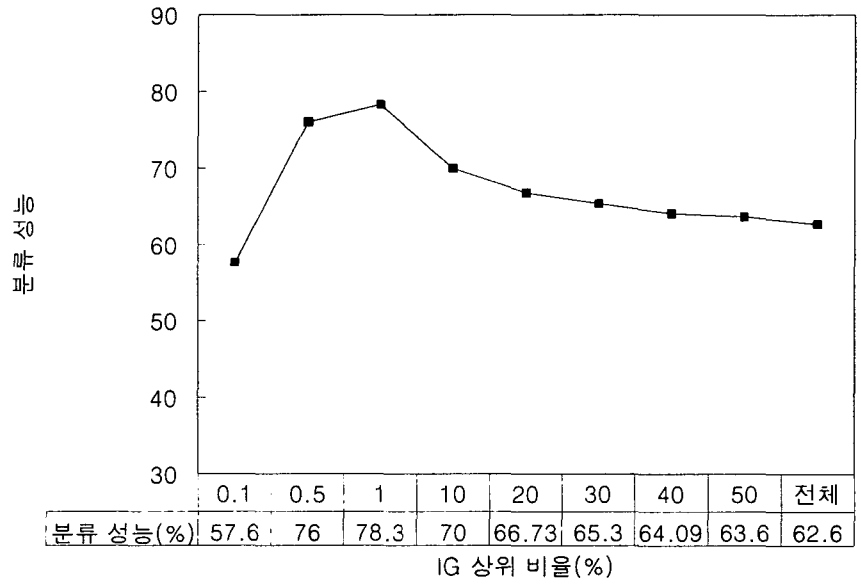

그림 1 특징수에 따른 분류성능의 비교

그림 2의 실험 결과는 각각의 특징 선택 기법으로 상위 1%의 단어를 선택하고, 특징을 구성하는 값을 Tfldf로 표현하였을 때, 각 특징 선택 기법으로 선택된 단어로 구성된 패턴의 분류 성능을 보여준다. 실험 결과에서 나타난 것 처럼 네 가지 특징 선택 기법 중 IG가 가장 높은 성능을 보였고, Gini index와 Relief-F가 거의 유사한 분류 성능을 보였다. 그러나 IG나 Gini index에 비하여 Relief-F는 복잡한 계산을 수행해야 하므로, IG기법이 특징 선택의 특징 선택을 하기 위해서 사용하는 방법으로 가장 적합하다고 사료된다.

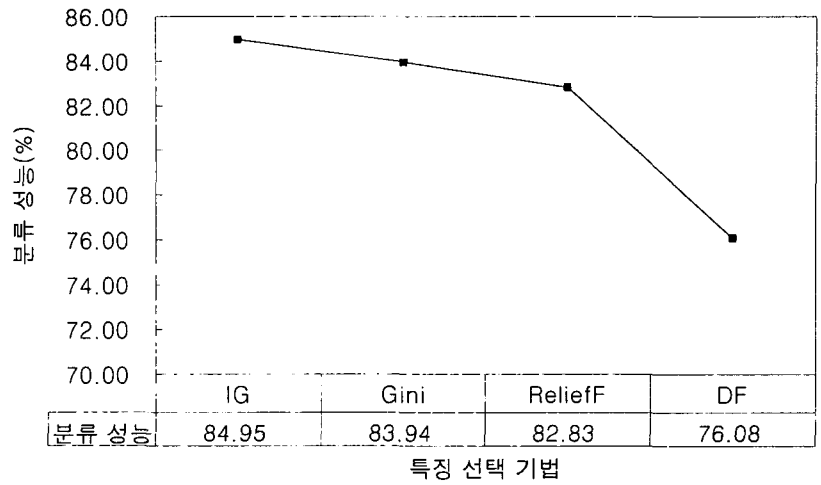

그림 2 특징선택기법에 따른 분류성능 비교

5. 결론

본 논문에서는 네 가지 특징 선택 기법인 IG, Gini index, Relief-F, DF를 이용하여 전체 문서에 포함된 고유 단어 중 일부를 선택하고, 이것을 특징의 개수로 하는 패턴의 분류 성능을 비교하였다.

실험 결과 IG를 특징 선택 기법으로 하여 고유 단어 중 1%만을 선택하고, 특징의 값을 TfIdf로 표현하여 패턴을 구성할 때, 가장 좋은 분류 성능을 보였다. 이것은 전체 고유 단어를 사용할 때, 패턴의 길이가 19,588개인데 반하여, 그 중 1%인 195개만을 패턴의 길이로 사용하기 때문에 본 논문에서 사용한 NN분류기 이 외의 다른 기계학습 기법을 적용하는 것이 가능하다고 사료된다.

참고 문헌

[1] J. R. Quinlan, Induction of Decision Tree, Machine Learning, 1, 81-106, 1986.

[2] Breiman L., Friedman J. H., Olshen R. A., Stone C. J., Classification and Regression Tree, Wadsforth International Group, 1984.

[3] I. Kononenko, Estimating Attributes: Analysis and Extensions of REFIEF, ECML94, pp171-182, 1994.

[4] Y. Yang, Jan O. Pedersen, A Comparative Study on Feature Selection in Text Categorization, Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), 1997.

[5] Dietrich Wettschereck, Weighted kNN vs. Majority kNN A recommendation, GNRCIT, 1995.

[6] David D. Lewis, Reuters-21578 text categorization test collection Distribution 1.0, README file(v1.2), AT&T Labs Research, 26 September, 1997

[7] Willam B. Frakes, Ricardo Baeza-Yates, Information Retrieval, Prentice Hall, 1992