문서분류에서의 SVM 및 나이브베이지안, EM알고리즘의 특성 비교

A Study on Comparison with SVM, EM, and Naivebayes Algorithm

저자
(Authors)

김병주
Kim Byung Joo

출처
(Source)

대한전자공학회 학술대회 , 2009.7, 683-684(2 pages)

발행처
(Publisher)

대한전자공학회
The Institute of Electronics and Information Engineers

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE02336080

APA Style

김병주 (2009). 문서분류에서의 SVM 및 나이브베이지안, EM알고리즘의 특성 비교. 대한전자공학회 학술대회, 683-684

문서분류에서의 SVM 및 나이브베이지안, EM알고리즘의 특성 비교
(A Study on Comparison with SVM, EM, and Naivebayes Algorithm)

김 병 주
(Kim Byung Joo)

Ⅰ. 서 론

현대 사회에 와서 다양한 형태의 많은 문서들이 문서들을 일괄적으로 분류 및 검색 할 수 있는 기술이 필요하게 되었다. 본 논문에서는 이러한 분야에 응용되는 알고리즘 중에서 SVM(Support Vector Machine)과 EM(Expectation Maximization), Naive Bayes에 대해서 살펴보고자 한다. SVM은 최대 분류 마진을 통한 경계면 설정이라는 특성을 가지고 있으며 이와는 반대로 확률을 기반으로 하는 EM과 Naive Bayes 분류기가 있다 이 3가지 알고리즘의 특성을 비교해 보고자 한다. 본 논문에서는 이러한 알고리즘 별 특성과 SVM에서 kernel의 종류에 따른 분류 성능과 특징추출의 종류에 따른 성능 향상을 살펴본다.

Ⅱ. 본 론

1.1 문서분류의 개념

문서분류는 학습데이터를 통해서 문서의 클래스를 학습하고 이를 바탕으로 분류모델을 형성한다. 이러한 분류모델 생성 알고리즘으로는 Naive Bayes, EM(Expectation Maximization), SVM(Support Vector Machine)이 있다. 분류알고리즘에 의해 학습되기 전에 문서들은 벡터 공간에서의 계산 효율을 높이기 위해서 전처리 과정을 거쳐야 한다.

1.2 관련연구

1.2.1 자질추출방식 : 분류모델형성에 불필요한 단어를 제거

1) TF-IDF (Term-frequency Inverse Document frequency)

TF-IDF=(term frequency)X(inverse document frequency) [식. 1]

2) 정보 획득량(Information Gain)

$$G(T)=-\sum_{i=1}^{k}Pr(C_i)\log Pr(C_i)+Pr(T)\sum_{i=1}^{k}Pr(C_i|T)\log Pr(C_i|T)$$

$$+Pr(\overline{T})\sum_{i=1}^{k}Pr(C_i|\overline{T})\log Pr(C_i|\overline{T}) \quad [식. 2]$$

1.2.2 분류 알고리즘

1) 나이브 베이즈(Naive Bayes Algorithm)

k 개의 범주 $c_1, c_2, \ldots, c_k$를 범주 집합이라고 할 때 새로운 문서 D가 범주 C_i 에서 나타날 확률 $P(C_i|D)$ 중 가장 확률에 문서의 범주 부여

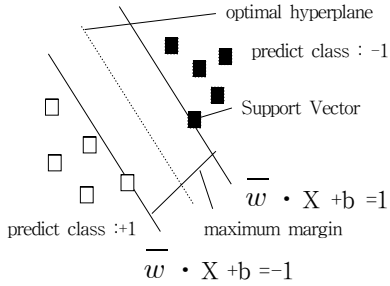
$$\max_{C_i} P(C_i|D)=\max_{C_i}\frac{P(C)P(D|C_i)}{P(C)} \quad [식. 3]$$

2) EM(Expectation Maximization)

EM알고리즘은 기본 분류 알고리즘을 바탕으로 문서를 분류한 후 분류된 문서를 다시 학습하는 방식이다. 이 학습방식은 분류 모델이 일정한 정확도를 나타낼 때 까지 반복된다. 기본 분류모델로서 나이브 베이즈 분류기가 사용된 실험[1]이 있다.

3)SVM(Support Vector Machine)

벡터 공간에서의 최대 분류 마진을 찾아내는 알고리즘으로서 최근에 각광받고 있는 학습 알고리즘이다.



[그림.1 두 클래스를 나누는 경계면]

[그림. 1]에서의 maximum margin은 다음과 같다.

$$\text{Maximum margin}=\frac{2}{\sqrt{\overline{w}\cdot\overline{w}}} \quad [식. 4]$$

SVM은 벡터공간에서 2진 분류를 시행하는데 각 학습 데이터(X_i)의 벡터가 올바른 클래스(Y_i)에 존재하는 제한조건이 있다.

제한조건 : $\overline{w}\cdot X_k + b = 1$ if $Y_k=1$

$\overline{w}\cdot X_k + b = -1$ if $Y_k=-1$ [식. 5]

제한 조건과 최대 마진(Maximum Margin)을 구하는 식을 결합하기 위해서 라그랑지(Largrangian Multiplier)공식을 적용한다.

$$\min_{\overline{w},b} L=\frac{1}{2}\overline{w}\cdot\overline{w}-\sum_{k=1}^{n}\lambda_k(y_k(\overline{w}\cdot X_k+b)) \quad [식. 6]$$

각 $\overline{w}$ 와 b에 대해서 미분하게 되면 0이 아닌 λ_k와 support vector인 X_k를 통해서 최대마진을 가지는 $\overline{w}$ 를 구한다. 벡터공간에서의 2진 분류가 어려운 경우 고차원공간으로의 사상을 통해서 2진 분류가 가능하게 만든다. 고차원 벡터 공간으로의 사상을 해주는 역할을 하는 함수를 커널(kernel)함수라고 한다.

기본 커널 함수는 다음과 같다.

Polynomial : $K(X_i, X_j) = (\Upsilon X_i^{T} X_j + r)^d, \Upsilon > 0$ [식. 6]

Linear: $K(X_i, X_j) = X_i^{T} X_j$ [식. 7]

Radial basis function : $K(X_i, X_j) = \exp(-\Upsilon |X_i - X_j|)^2, \Upsilon > 0$ [식. 8]

Sigmoid: $K(X_i, X_j) = \tanh(\Upsilon X_i^T X_j + r)$ [식. 9]

본 논문에서는 svm의 마진과 확률을 근거로 하는 EM(Expectation maximazation)과 Naive Bayes의 특성을 비교해보고 svm에서 학습 문서가 적을 경우와 많을 경우에, 분류 평면(classification hyperplane)의 형성에 대해서 살펴보기로 한다. 특징 추출(feature extraction)에서 tf-idf와 information gain을 적용했을 경우의 성능 향상을 실험하였다, 그리고 kernel의 종류에 따른 정확도를 실험한다.

Ⅲ. 실 험

컴퓨터와 관련된 5개 분야의 인터넷 문서 500개를 가지고 실험을 하였다. 문서 분류 알고리즘으로 EM, Naive Bayes, SVM을 적용하였을 경우 문서의 수가 적을 경우에는 정확도가 대부분의 경우에 정확도가 낮게 나온다. 학습 문서의 수가 많아짐에 따라서 EM과 Naive Bayes 알고리즘은 정확도가 올라감을 실험을 통해서 알 수 있었다. 학습 문서의 수가 적을 경우에 SVM의 경우 EM과 Naive Bayes에 비해서 정확도가 높게 나왔다. 또한 SVM은 학습 문서가 많아져도 정확도가 크게 향상되지 않았다. 이는 SVM의 분류 경계면을 구할 때, 식 (6)을 라그랑지 공식을 적용하여 dual problem 형식으로 변할 경우 $w=\sum_{i=1}^{l} \lambda_k y_k x_k$가 되어서 hyper plane에 근접한 $\lambda_k>0$을 만족하는 소수의 x_k(support vector)만이 학습 데이터로 학습되기 때문이다.

	학습문서(20%)	학습문서(50%)	학습문서(80%)
SVM	87%	93%	95%
EM	67%	77%	91%
naive bayes	65%	78%	89%

<표. 1 분류 알고리즘 별 정확도>

위 실험을 통해서 SVM은 학습 데이터의 양에 크게 의지하지 않고도 높은 정확도를 얻는 것을 볼 수 있다. 이 결과는 SVM이 단지 hyper plane의 support vector만을 이용하여 최적의 분류 경계면을 형성하기 때문에 학습 데이터가 늘어나도 분류 경계면이 크게 변하지 않는 다는 것을 보여 준다. 다음은 SVM의 kernel 함수에 따른 정확도의 차이를 실험하여 보았다.

	linear kernel	sigmoid kernel
Test 문서의 20%	94.1%	57.0%
Test 문서의 50%	97.6%	65.0%
Test 문서의 80%	99.0%	68.5%

<표. 2 SVM의 kernel 함수별 정확도>

일반적으로 sigmoid kernel 함수 [식. 9]이 linear kernel 함수 [식. 7]보다 일반적으로 우수한 것으로 알려져 있으나 kernel 함수를 통해서 고차원 공간으로의 사상이 필요 없이 linear하게 학습 벡터 공간이 분류가 가능한 경우에는 linear kernel 이 높은 성능을 보이게 된다. sigmoid kernel의 경우는 오히려 고차원 공간으로의 사상을 통해서 더 낮은 분류 성능을 보이게 되어서 문서분류의 경우는 선형 분리의 확률이 높다는 것을 알 수 있다. polynomial kernel [식. 6]과 radial basis kernel [식. 8]의 경우는 문서 분류의 성능이 위 실험의 두 kernel 함수에 비해서 낮게 나왔다. 이를 바탕으로 문서 분류의 경우는 복잡한 커널 함수의 적용하는 것보다 최적의 경계면을 형성하는 경계면을 사용하여 분류를 시행하는 것이 효과적이다. 다음의 표는 단어를 추출할 경우, 사용하는 feature extraction 함수에 따른 성능 향상 실험을 보여준다.

	raw	tf-idf	info-gain
svm precision	60.1%	67.9%	61.5%

[표. 3 feature extraction에 따른 정확도]

[표. 3]의 실험 결과는 단어 추출에 있어서 inverse document frequency를 사용할 경우에 일반적인 stemming 과정을 통한 특징을 사용하는 경우보다 더 높은 정확도를 보여준다. information gain을 통해서 특징(단어)추출을 하였을 경우 특징추출을 하지 않은 경우와 큰 차이가 없는 정확도를 보여준다. 이는 SVM에서 확률을 근거로 하는 information gain 특징 추출은 support vector만이 분류 경계면을 형성하기 때문에 성능향상에 크게 도움을 주지 않는 것을 보여준다.

Ⅳ. 결 론

SVM(Support Vector Machine)이 적은 학습 데이터에도 불구하고 높은 분류 정확도를 보여주는 이유는 support vector를 이용한 분류 경계면(hyper plane)을 형성하기 때문이다. 확률 분포를 분류근거로 삼는 EM(expectation maximization)과 Naive Bayes는 학습 데이터의 양이 늘어남에 따라서 정확도가 올라가는 사실은 이 두 분류기가 학습데이터가 적을 경우에는 정확한 분류를 시행하기 어렵다는 것을 나타내 준다. SVM을 문서 분류에 적용하였을 경우, 한정된 분야의 경우는 학습 벡터 공간을 선형으로 분리의 확률이 높기 때문에 고차원 공간으로의 사상은 불필요하다는 것을 실험을 통해서 알 수 있었다. 앞으로의 실험은 SVM의 특성을 알아보고 분류 성능 향상을 위해 인터넷 공간의 계층 구조의 문서 분류와 연관성이 없는 분야 간의 문서 분류에 대해서 실험할 필요가 있다.

참 고 문 헌

[1] Semi-Supervised Text Classification Using EM Kamal Nigam, Andrew McCallum , Tom M. Mitchell

[2] A Study on Sigmoid Kernels for SVM and the Training of non PSD Kernels by SMO-type Methods Hsuan-Tien Lin and Chih-Jen Lin Department of Computer Science and information Engineering National Taiwan University

[3] Efficient SVM regression Training with SMO Gary William Flake, Steve Lawrence, NEC Research institute

[4] Learning to Classify Text using Support Vector Machines, thorsten johochim Theory, and Algorithms Cornell University Department of Computer Science

[5] McCallum, A., "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering" ,1996