문서 분류에서의 SVM 오류 감소를 위한 하이브리드 방법

이준석° 김상수 박성배 이상조
경북대학교 컴퓨터공학과
{jslee°, sskim}@sejong.knu.ac.kr {sbpark, sjlee}@knu.ac.kr

Hybrid Approach to SVM Error Reduction in Document Classification

Jun-Seok Lee° Sang-Soo Kim Seong-Bae Park Sang-jo Lee
Dept. of Computer Engineering, Kyungpook National University

요 약

본 논문에서는 문서 분류(document classification) 성능을 높이기 위해 다음과 같은 방법을 제안한다. 먼저 패턴 분류 문제에 있어서 우수한 성능을 보이는 SVM(Support Vector Machine)을 사용하여 분류 하고, 마진을 만족하는 데이터를 다시 k-NN 으로 분류를 한다. 단순히 SVM만을 사용한것보다 k-NN을 함께 사용한것이 더 높은 성능을 보였다.

1. 서 론

인터넷의 발전으로 인하여 수많은 정보들이 폭발적으로 증가하고 있고, 이러한 정보들 중에서 사용자가 원하는 정보를 얻기가 매우 어려워 졌다. 따라서 사용자가 필요한 정보를 획득하기 위해서 정보를 담고 있는 문서를 자동 문서 분류 시스템(Automatic Text Classification System)을 사용해서 사용자에게 제공하는 방법이 대두되었다.

자동 문서 분류는 미리 정의된 각종 분류기법을 사용하여 자동으로 분류되지 않은 문서를 미리 정의된 클래스들로 나누는 것을 말한다. 일련의 연구 결과에 의하면 분류체계를 도입함으로써 검색엔진만 사용하는 검색에 비해 소요시간이 약 50%정도 감소되었다고 보고되고 있다[1].

자동 문서 분류 시스템은 크게 그 문서를 잘 표현하는 자질(feature)를 추출하는 과정과 추출된 자질을 이용하여 분류하는 부분으로 나누어 볼 수 있다. 문서를 잘 표현하는 자질을 추출하는 과정은 통계학적인 방법을 사용하는데, 그 방법들로는 코사인 유사도(cosine similarity), 상관계수(correlation), 정보 이득(information gain), 교차 엔트로피(cross entropy), 상호 정보(mutual information), odds-ratio등이 있다. 추출된 자질을 이용하여 분류하는 분류 과정은 주로 기계 학습 방법을 사용하는데, 주로 나이브 베이스 분류자(naive Bayes' classifier)가 많이 이용되며 이밖에도 SVM(support vector machine), 선형분류자(linear classifier), k-NN 등을 이용하고, 또한 학습의 성능을 높이기 위하여 부스팅(boosting) 기법이나 레이블이 없는 문서(unlabeled documents)를 이용하기도 한다.

본 논문에서는 자질 선택 방법으로 정보 이득(Information Gain)을 사용하고, 분류 학습 기법은 이진 분류에서 우수한 성능을 보이는 SVM을 사용했고, 학습 및 실험에 사용된 문서 집합(corpus)은 REUTERS-21578을 사용하였다. 학습 과정은 먼저 SVM을 통하여 분류하고 SVM의 마진이 특정 부분에 속하면 다시 k-NN을 사용하여 분류를 했다. 실험은 단순 SVM만 사용한 것과 k-NN을 함께 사용한 것을 비교 실험 하였다.

본 논문은 2장에서는 관련 연구에 대하여 알아보고, 3장에서는 본 논문의 시스템에 대하여 설명하고, 4장에서는 실험, 5장에서 결론을 살펴본다.

2. 관련연구

본 장에서는 자질선택 방법과 SVM 과 k-NN에 대하여 설명한다.

2.1 정보 이득을 이용한 자질선택 방법

특정 범주에 연관된 단어들을 가지고 임의자질 선택 하는 방법에는 여러 가지가 있지만 정보 이득을 사용했다. 정보 이득은 클래스를 결정할 때, 문서에서 어떤 단어의 유무를 통해 단어와 부류간의 관계를 알아낸다. 전체 문서에서 부류들의 집합을 $\{c_i\}_{i=1}^{m}$ 이라고 할 때, 단어 t의 정보 이득은 식 1과 같이 정의 된다[2].

$$G(t) = -\sum_{i=1}^{m} \Pr(c_i)\log Pr(c_i) + \Pr(t)\sum_{i=1}^{m} \Pr(c_i|t)\log Pr(c_i|t) + \Pr(\bar{t})\sum_{i=1}^{m} \Pr(c_i|\bar{t})\log Pr(c_i|\bar{t}) \quad (1)$$

식 1에 사용된 속성이라는 것은 특정한 범주에 속하는 단어를 의미한다. 계산된 속성들의 정보이득 값을 통해 단어가 갖는 분류 확률의 순위를 매길 수 있으며, 정보 이득 값에 근거해서 단어들을 선택할 수 있게 된다.

2.2 SVM(Support Vector Machine)

SVM은 분류(classification)와 회귀(regression)에 응용할 수 있는 지도학습(supervised learning)의 일종이다. SVM 은 Vapnik[2]에 의해 서 1995년 이진문제를 해결하기 위해서 제안된 알고리즘이다. SVM은 선형적으로 분리할 수 있는 학습집단에 대해서 최대마진 분류기를 구축하는 선형 SVM과 선형적으로 분리할 수 없는 경우에 커널 함수에 의해 만들어지는 비선형 결정함수를 이용하는 최적의 초평면(hyperplane)을 구축하는 비선형 SVM으로 분류된다.

SVM은 식 2와 같이 학습 데이터를 두 클래스로 정확하게 분류하는 최적의 초평면을 찾는다.

$$(w \cdot x) + b = 0 \tag{2}$$

$$w \in R^n \ , \ b \in R^n \tag{3}$$

이를 위해 초평면과 가장 인접한 점과의 거리(margin)가 최대가 되도록 초평면을 학습한다. 이 거리 (d)는 식 4과 같이 나타낼 수 있다.

$$(w \cdot x) + b = \pm 1 \tag{3}$$

$$d = 2 / \| w \| \tag{4}$$

SVM이 주목받는 이유는 첫째, 명백한 이론적 근거에 기반하므로 결과 해석이 용이하고, 둘째, 실제 응용에 있어서 인공신경망 수준의 높은 성과를 내고, 셋째, 적은 학습자료만으로 신속하게 분별학습을 수행할수 있기 때문이다. 이렇기 때문에 우리는 SVM 도구로서 SVMlight를 사용하였다[3].

2.3 k-NN(k-nearest neighborhood)

k-NN 분류 방법은 그 방법의 간단성과 직관성 때문에 여러 분야에서 널리 이용되고 있다. 학습자료 $X_1, \cdots, X_m$은 모집단 X에서 그리고 $Y_1, \cdots, Y_m$은 모집단 Y에서 추출된 표본이고 새로운 관측값이 Z 일때 k-NN 분류자 $\bar{\theta}(Z)$는 식 5와 같이 정의 된다[4].

$$\bar{\theta}(Z) = \text{type } X \text{ if number of } X_i \text{'s in } N_{k(z)} \geq k/2$$

$$= \text{type } Y \text{ otherwise} \tag{5}$$

여기에서 $N_k(z)$는 Z 의 k-NN로서 $X_1, \cdots, X_m$ 와 $Y_1, \cdots, Y_m$중 Z와의 거리가 제일 가까운 것부터 k 번째로 가까운 학습자료를 포함한 것이다. 즉, 새로 관측된 Z 의 type은 그것의 k-근방에 포함된 학습자료 중 더 많이 포함된 type으로 예측된다.

3. 문서 분류 시스템

문서 자동분류는 각종 분류기법을 이용하여 자동으로 새로운 문서를 미리 정의된 부류들로 나누는 것을 말한다. 그 과정은 그림 1에서와 같이 크게 자질 추출 단계, 문서 분류 단계로 구분되고, 학습 및 실험에 사용된 문서는 스테밍, 불용어 처리 등을 거친 단어들로 구성하였다.

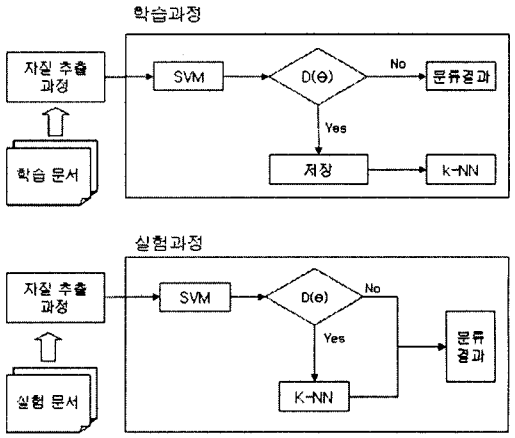

그림 1 자동 문서 분류 시스템 구성도

자질 추출 단계에서는 학습 및 실험 문서들은 전체 26040종류의 단어들로 구성되어 있고, 이들 단어들을 벡터로 표현하면 너무 많은 차원을 가지게 된다. 벡터의 차원이 크면 학습의 속도가 늦어지고, 문서분류 성능 또한 나빠지게 되는데, 이러한 문제를 극복하기 위하여 차원을 줄일 필요가 있다. 본 논문에서는 정보이득 자질선택 방법을 사용하여 단어와 문서간의 정보이득을 계산하여 표 1과 같이 나타 내었다. 표 1은 문서의 단어 정보이득 값이 가장 높은 10%부터 100%까지 변화 시켜가면서 정확률 값을 나타내었다. 실험한 결과 55%(14000차원)일 때 가장 높은 값을 가지는 것을 알 수 있다.

표 1. 자질(%)수 에 따른 정확률

자질(%)수	10	20	···	55	···	90	100
정확률	72.05	72.26	···	72.59	···	72.42	72.40

문서 분류 과정에서는 SVM과 k-NN을 사용하였다. SVM은 이진 분류 문제에서 장점을 가지고, k-NN은 예제 기반의 학습 방법으로 예로 주어진 자료에서 최적의 해를 구한다. 그러나 k-NN은 주어진 예가 많으면 사용하기에 부적합하다. 그래서 본 논문에서는 문서 분류 방법으로 거리함수(Distance function) D(θ)를 통하여 일반적인 분류를 처리하는 SVM과 예외 처리를 위한 k-NN 모델로 구성하였다. D(θ)는 입력되는 문서 벡터가 SVM을 통하여 나오는 값(마진)을 참조하여 실험적으로 구하였다.

4. 실험 및 분석

학습 및 실험에 사용된 문서 집합(corpus)은 독일 Reuter 신문의 웹 기사인 REUTERS-21578을 사용하였고, 표2에 나타난 것과 같이 총 19043개의 문서로 구성되어 있다. 실험은 이중 13186의 문서를 학습에 사용하고, 5857의 문서를 실험에 사용하였다.

REUTERS-21578은 총 135의 토픽으로 분류되어 있고, 각각의 문서들은 토픽이 분류되어있지 않거나, 중복이 허용된다.

본 논문에서는 5개의 범주 중에서 EARN범주 부분과 그렇지 않은 부분으로 이진 분류로 실험을 수행 하였다. 즉 EARN 분류는 포지티브(positive), 그렇지 않은 분류는 네거티브(negative)로 보았다.

표 2. 실험에 사용된 전체 문서의 수

	학습 문서 개수	테스트 문서 개수
positive	2758	1018
negative	10428	4839
총 문서 개수	13186	5857

SVM에서 잘못 분류한 문서를 판단하는 거리 함수 D(θ) 에서 k-NN으로 실험 할 문서의 수는 표3과 같다.

표 3. k-NN 에 사용된 문서 수

	학습 문서 개수	테스트 문서 개수
positive	406	134
negative	607	419
총 문서 개수	1013	625

그림 3는 SVM으로 분류했을 때의 잘못 분류한 것을 나타낸 것이다. 표에서 가로축은 문서수를 세로축은 SVM에서 나온 마진 값을 나타낸 것이다. 그림 2에서 보면 negative 인데 positive로 분류된 문서를 k-NN을 사용하여 바로 잡았다.

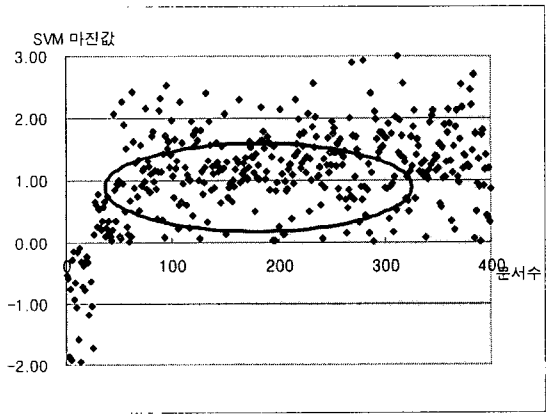

그림 2. 잘못 분류한 집합

그림 3에서는 k-NN의 k 수를 변화 시켜 가면서 실험한 결과를 나타내었다. 여기서는 k=563일때 실험 결과가 가장 높게 나왔다. 그래서 k-NN에서 k= 563을 사용한다.

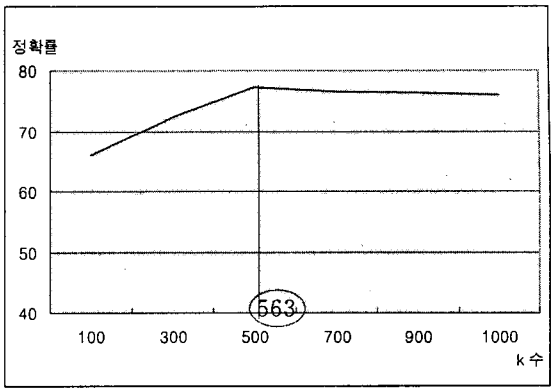

그림 3. k 수 변화

표 4는 자질의 수를 14000(55%)개로 같게하고, SVM만 실험한 결과와 k-NN을 함께 실험한 결과 표이다.

표 4. SVM 과 k-NN을 함께 사용한 실험

	SVM	SVM + k-NN
정확률	72.59	78.14

5. 결론 및 향후 연구과제

본 연구에서는 단순히 SVM만 사용해서 분류하는것 보다 k-NN을 사용해서 한번더 분류해 줌으로서 좀더 나은 성능을 보였다. SVM만 사용해서 분류한 재현률은 거의 차이가 없었고, 정확률은 더 높은 성능을 보였다

본 연구에서는 정보이득만을 가지고 자질을 선택 하였는데 자질 선택 방법을 다양하게 하는 연구가 필요하다.

참고문헌

[1] Chen, Hao and Dumais, Susan, Bringing Order to the Web: Automatically Categorizing Search Results, Proceedings of CHI2000, pp. 145-152, 2000.

[2] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization, " *Proc. of 14th Int. Conf. on Machine Learning*, pp.412~420, 1997.

[3] http://svmlight.joachims.org.

[4] Mitchell, T.M (1997). *Machine Learning*. The MCGraw-Hill Companies, Inc.