SVM 분류 알고리즘을 이용한 스팸메일 필터링

민도식[0], 송무희, 손기준, 이상조
경북대학교 컴퓨터공학과
{sahel[0], muhee, kjson}@sejong.knu.ac.kr sjlee@knu.ac.kr

Spam-mail Filtering Using SVM Classifier

Do-Sik Min[0] Mu-Hee Song Ki-Jun Son Sang-Jo Lee
Dept. of Computer Engineering, Kyungpook National University

요 약

전자우편은 기존 우편 기능을 대체하는 대표적인 정보 전달 수단으로 자리 잡고 있다. 전자메일 사용자의 증가에 따라 많은 기업들은 전자 메일을 통해 광고를 하게 되었다. 이에 따라 전자메일 사용자들은 인터넷 상에 개인 전자메일 주소가 노출됨으로 많은 스팸메일을 수신하게 되는데, 이것은 전자메일 사용자에게 많은 부담이 되고 있다. 본 논문은 전자우편 문서내의 단어들을 대상으로 통계적 방법의 SVM을 이용하여 스팸메일을 필터링 하였으며, 학습 단계에서 단어 자질공간의 축소를 위해 DF값 변화에 따른 학습을 통하여 분류의 성능을 비교하였다. SVM의 성능 평가를 위해 확률적 방법의 나이브 베이지안과 벡터 모델을 이용한 분류기와 성능을 비교함으로써 SVM 방법이 우수한 성능을 보임을 검증하였다.

1. 서 론

전자우편 문서 사용자들은 자신의 관심사나 개인적이고 중요한 문서를 메일을 통해서 주고 받는다. 하지만 전자 우편 문서는 많은 광고성 스팸메일을 포함하고 있고, 이런 스팸 메일을 걸러 줄 수 있는 필터링에 대한 관심이 높아 졌다.

일반적으로 스팸메일 필터링은 메일을 받아서 스팸인지 아닌지, 두 가지 범주로 분류하는 문제로 본다. 전자 우편 문서는 일반 웹 문서와는 다르게 비형식적인 문장 구조를 가지고 있기 때문에 문서를 분류하는 것에 많은 어려움이 있다. 이러한 문제들은 기존의 규칙 기반을 이용한 방법만으로는 해결하기 힘들다. 따라서 필터링 하는데 있어서 전자문서 분류에 많이 이용되는 확률적인 방법이나 통계적인 방법을 이용할 필요가 있다.

SVM(Support Vector Machine)은 통계적 학습이론에 기반한 방법으로 경험적 리스크를 최소로 하는 것이 아닌, 구조적 리스크를 최소로 하는 이진 패턴 분리를 위한 알고리즘이다[1]. 전자문서 분류에 대한 연구에 많이 이용되는 확률적인 방법으로는 나이브 베이지안이 있으며, 많은 학습 문서를 필요로 한다. SVM은 전자문서 분류에서 적은 학습문서에서도 높은 정확율을 보이고 있고, 이와 같은 이진 분류 문제에 적합하다.

본 논문은 전자우편 문서의 제목과 본문에 나오는 단어를 대상으로 자질을 선택하였고, 최근에 문서 분류에서 높은 성능을 보이고 있는 분류 방법인 SVM을 사용하였다. 또한 SVM 사용에서는 학습단계에 DF값 변화에 따른 자질공간의 축소를 통해 나은 성능을 보이는 DF값을 임계값으로 하여 나이브 베이지안과 벡터 모델을 이용한 분류기와 성능을 비교 분석 하였다.

본 논문의 구성은 다음과 같다. 2절에서는 관련연구에 대해서 살펴보고, 3절에서는 SVM을 이용한 스팸메일 필터링에 대해 설명하며, 4절에서는 실험 및 평가를 하고, 마지막으로 5절에서 결론과 향후과제에 대해 기술한다.

2. 관련연구

기존의 전자우편 문서 분류에 대한 연구는 규칙기반을 이용한 전자우편 문서의 반구조성을 적용한 Cohen[2]의 RIPPER 시스템이 있었다. 이와 같은 규칙 기반의 시스템이 나이브 베이지안을 이용한 확률 모델과 비교 실험에서 스패머들에 의해 광고의 내용이 자주 변동하기 때문에 키워드나 규칙을 이용하는 것에 대응하기가 힘들며, 스팸메일 분류에 적당하지 않다는 연구 결과가 보고되었다[3].

또한 일반문서 분류에서 SVM은 나이브 베이지안 보다 좋은 성능을 보였다. Thorsten Joachims[4]는 Naive Bayes, Rocchio, k-neareat neighbors, C4.5, SVM의 다섯 가지 알고리즘을 비교했다. 실험 결과 SVM과 나이브 베이즈를 제외하고 최적의 자질 수는 전체 자질 수보다 작았다. SVM의 경우 모든 자질을 사용함으로써 다른 분류방법을 사용하는 것보다 좋은 수행을 얻을 수 있었다. 국내에서 나이브 베이지안과 메시지 규칙을 이용한 스팸메일 필터링에 관한 연구가 있었다[8].

2.1 SVM 분류기

SVM은 선형적으로 분리할 수 있는 학습집단에 대해서 최대 마진 분류기를 구축하는 선형 SVM과 선형적으로 분리할 수 없는 경우에 커널 함수에 의해 만들어지는 비선형 결정함수를 이용하는 최적의 초평면을 구축하는 비선형 SVM으로 분류된다. 선형 SVM은 데이터를 +1 와 -1클래스의 두개의 집합으로 완전하게 분리시킬 수 있는 결정면이 존재한다는 것이며, 이것은 다음의 수식(1)로 나타낼 수 있다[5].

$$\exists w, b \text{ 일 때 } \begin{cases} w \cdot x_i + b > 0, & y_i = +1 \\ w \cdot x_i + b < 0, & y_i = -1 \end{cases} \quad (1)$$

w는 가중치벡터, x는 입력벡터, b는 기준치이며, w와 b는 학습 데이터로부터 학습된다. 학습 문서 집합을 $D = \{(x_i, y_i)\}$라고 하면, 입력데이터 x_i가 범주에 속하면 y_i는 +1의 값을 갖고, 속하지 않으면 -1의 값을 갖는다. 결국 SVM은 최적의 w와 b을 찾는 문제이다.

2.2 나이브 베이지안 분류기 및 벡터 모델

베이지안 분류기는 전자우편 문서에 나타난 단어들의 분포는 서로 독립임을 가정하며, 단어가 나타날 확률은 문서내에서 단어의 위치와도 독립적이라고 가정한다. 베이지안 분류기는 범주가 출현할 사전 확률을 기반으로 하여 스팸과 논스팸으로 나누어진 범주가 문서에 할당될 확률과, 특정 단어가 전자우편 문서에서 발생할 조건 확률을 계산 분류하려는 문서 D에 단어 $Term_i$가 출현한 경우 전자우편 문서가 범주 C_j에 분류될 확률을 아래 수식(2)과 같이 계산된다[6].

$$P(C_j \mid D) = \underset{i \in position}{\arg\max} \; P(C_j) \cdot \sum_{i=1} P(Term_i \mid C_j) \quad (2)$$

벡터 모델에서 두 문서 사이의 유사도 측정에는 코사인 유사도를 이용한다. 문서 x와 y에 대한 유사도는 아래의 수식(3)과 같다[7].

$$Sim(x, y) = \frac{\sum (x_i \cdot y_i)}{\sqrt{\sum_{i=1} (x_i)^2} \cdot \sqrt{\sum_{i=1} (y_i)^2}} \quad (3)$$

3. SVM을 이용한 스팸메일 필터링

통신상에서 의사소통을 목적으로 하는 전자우편 문서는 개인적인 내용뿐만 아니라, 광고, 서비스, 제품 판매 등의 특정 목적에 이용되고 있고, 이렇게 사용자의 의사와는 관계 없이 메일을 통하여 사용자에게 전해지는 불필요한 정보를 스팸이라 한다.

스팸 문서의 특징은 기존의 일반 문서와는 다르게 일반적인 문장 구성을 이루지 않고, 특수문자, 약어, 속어 및 많은 이미지 데이터 등을 포함한 형태의 교묘한 방법으로 사용자에게 스팸메일을 전달한다. 이러한 특성은 스팸메일 필터링에 정확율을 낮추게 하는 문제점이 되고 있다.

기존의 메일 필터링 시스템이 가지고 있는 단순한 규칙 기반 필터링 시스템은 이러한 스팸메일을 필터링 하는 것이 힘들다. 간단한 예로, '광고'라는 문구가 들어 있는 스팸메일에 대한 규칙 기반 필터링에서는 특정 단어의 매칭을 통해 필터링하게 되겠지만, 지능화된 스패머(Spammer)들에 의해 제목이나 본문의 내용 중에 광고라는 문구를 '광_고', '광**고' 등의 변형된 형태에 대해서는 스팸 문서를 필터링 하기 힘들다는 것이다. 따라서 본 논문은 단어를 대상으로 자질을 선택하여 SVM을 통해 학습하고, 분류하는 방법을 제안하며, 단어를 대상으로 한 자질 선택에서 자질 공간 축소를 위해 DF값의 변화를 사용하였다. 그림 1은 본 논문에서 제안된 전자우편 문서 분류기의 시스템 구성이다.

분류기는 크게 세 가지 단계로 구성된다. 전처리 단계에서 전자우편 문서는 많은 속어나 약어를 사용하므로 학습 데이터의 크기가 많아진다. 이에 따라 전처리 과정으로서 불필요한 정

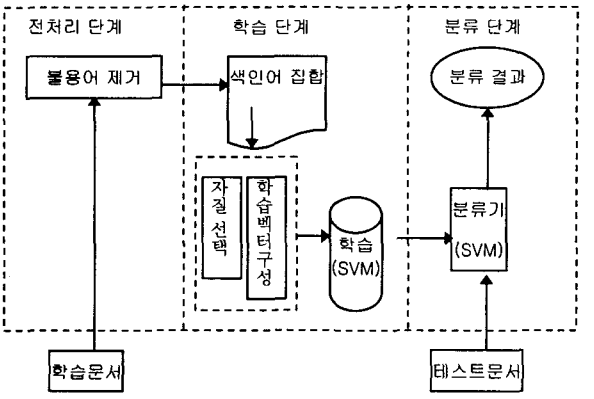

그림 1. 전자우편 문서 분류기의 시스템 구성

보를 제거하여 공간을 줄이기 위해 불용어 목록(stop list)을 사용한다. 이 단계에서 자질 공간을 줄이게 된다

학습 단계에서는 전처리 단계를 거친 텍스트에서 색인어를 추출하여 색인어 집합을 구성하고, 학습 데이터의 구성을 위해 색인어 집합을 이용하여 각각의 단어와 가중치의 쌍으로 이루어진 자질을 사용한다.

전자우편 문서는 일반 문서와는 다른 특징들을 가지고 있다. 문서 정보는 헤더 부분과 본문으로 나누어지고, 헤더 정보에는 메일의 제목, 보낸 사람, 보낸 사람의 도메인, 계정, 보낸 시간 등의 자질을 가지고 있고, 본문 부분은 메일의 내용이 포함되어져 있다. 본문에서 얻을 수 있는 자질은 단어에 대한 정보 이외에도 스팸이 가지고 있는 자질 중에 링크 정보 같은 것을 얻을 수 있다. 하지만, 자질들을 직접 학습 및 분류 방법에 적용하기가 힘들다.

본 논문에서는 헤더와 본문에 나오는 단어들을 사용하였으며, 자질 선택 방법으로 TF, TF*IDF를 사용한다. 분류 자질의 선정은 자질 공간의 축소를 의미한다. 즉 정보성이 낮거나 없는 단어를 효과적으로 배제하는 작업을 의미한다. 학습 데이터가 많아지면 분류에 더 나은 성능을 보일 수 있지만 비례적으로 자질의 수도 많아져 자질공간이 커지게 되고, 학습과 분류에 많은 시간을 필요로 하게 된다. 따라서 본 논문은 각각의 자질 선택 방법에 대해, DF값으로 자질 공간을 축소하기 위해 단어의 DF값에 따른 학습 데이터를 구성하였다. 각각의 학습데이터는 DF값을 1~4까지 주어 각각의 DF값 이상을 학습 데이터로 구성하여 학습하였다.

SVM 학습 데이터 구성을 위한 자질로서는 TF, TF*IDF 각각과 두 가지 자질을 모두 사용하여 세 가지 자질 벡터를 각각 학습하여 스팸메일 필터링에 사용한다. 나이브 베이지안 학습 데이터 구성을 위한 자질로서는 TF를 사용하였다.

각각의 학습된 문서는 선형 SVM와 나이브 베이지안 분류기를 통해 필터링하고, 벡터 모델은 코사인 유사도 공식을 이용하여 자질 벡터와 전자우편 문서 사이의 유사도로 스팸메일을 필터링 하였다.

4. 실험 및 평가

실험은 각각의 분류에 대해 동일한 학습 문서와 테스트 데이터를 사용하였다. 테스트 및 학습 문서는 실험을 위해 3개

월 분량의 메일을 사용하였으며, 학습문서 수는 666개의 문서를 대상으로 이루어진다. 실험 대상 문서 중 스팸 문서 수는 441개 이고 논스팸 문서 수는 225개로 구성 되어 있다. 그리고 분류기의 성능을 평가하기 위해 153개의 문서를 분류하는 실험을 하였다. 테스트 문서 중에 스팸 문서 수는 100개이고 논스팸 문서 수는 53개로 구성 되어 있다. 본 실험을 위해 SVM은 SVM Light[9]를 사용하였다.

표 1은 DF값의 변화에 따른 학습 데이터 구성을 통해 SVM 분류 실험 결과이다.

표 1. DF 변화에 따른 분류 결과

		DF(1)	DF(2)	DF(3)	DF(4)
TF	Precision (%)	90.83	91.67	91.67	93.46
	Recall (%)	99.00	99.00	99.00	100
TF*IDF	Precision (%)	91.18	91.43	91.67	92.59
	Recall (%)	93.00	96.00	99.00	100
TF & TF*IDF	Precision (%)	95.15	94.29	94.29	93.40
	Recall (%)	98.00	99.00	99.00	99.00

본 실험의 평가는 정확율과 재현율을 사용하여 성능을 평가하였다. 실험 결과는 DF 임계값 4에서 SVM을 이용해 학습하여 스팸메일을 분류할 때 나머지 다른 DF값 보다 높은 수치를 보였다. 자질공간을 줄이며 성능에 크게 영향을 주지 못하는 단어 자질들이 제거되어 다른 DF값 보다 나은 성능을 보였다.

본 실험에서는 DF값을 4까지만 주었고, 4이상의 DF값이 주어진다면 자질공간은 많이 축소 되겠지만, 중요한 정보를 가진 단어들도 제거될 것이다. 전체적으로 학습 데이터의 양이 늘어난다면 DF값을 이용하여 적당한 자질공간을 이용하여야 한다. 단어에 대한 자질로서 TF와 TF*IDF 그리고 두 가지 방법을 모두 이용한 실험에서 분류 결과는 비슷한 결과를 주었지만, 두 가지 자질을 모두 선택한 방법에서 좀더 나은 결과를 주었다.

표 2는 각각의 분류기에 대한 성능을 실험한 결과이다. SVM 분류기는 DF값 4일 때, 나이브 베이지안 분류기, 벡터모델을 이용한 방법과 비교하였다.

실험 결과, 표 2에서 TF와 TF*IDF 자질을 사용한 스팸메일 분류에 대해 SVM이 나이브 베이지안과 벡터 모델에 비해 정확율, 재현율에서 좋은 성능을 보였다. 그리고 각각의 자질을 달리한 SVM에서 큰 차이는 보이지 않았다. 적은 학습 문서와 테스트 데이터 였지만, 전체적으로 SVM은 좋은 성능을 보였다.

표 2. 분류기의 성능 비교

	SVM (TF)	SVM (TF*IDF)	SVM (TF&TF*IDF)	NB	VSM
Precision (%)	93.46	92.59	93.40	75.38	87.36
Recall (%)	100	100	99.00	98.00	76.00

전자우편 문서가 많은 데이터 양을 가지지 않기 때문에 통계적인 방법의 SVM을 사용하는 것이 나은 성능을 보였다.

5. 결론

스팸메일 필터링에서 나이브 베이지안을 이용한 분류와 벡터모델을 이용한 분류 보다 SVM을 이용한 분류가 좋은 결과를 주었고, SVM의 자질공간을 줄이기 위해 DF값을 임계치로 사용한 실험에서는 대부분 좋은 결과를 주었지만, DF값이 4일 때 더 나은 결과를 주었다.

이 실험에서는 선형 SVM만을 사용하였다. 그러나 SVM은 입력 데이터가 선형분리가 불가능할 경우 입력공간을 분리하는 비선형 결정면을 이용하게 된다. SVM은 커널 함수를 이용하여 고차원 자질공간에서 벡터로 변형한 후, 선형 경계선을 찾는 문제로 전환하는 비선형 SVM에서도 좋은 성능을 보인다고 한다[4]. 비선형 문제에 대한 성능도 비교할 필요가 있다.

스팸메일의 다양한 자질을 어떻게 학습 및 분류에 적용할 것인지에 대한 연구를 통해 어떤 자질이 스팸메일 필터링 시스템에 더 좋은 성능을 주는지에 대해서 향후 연구해 보고자 한다.

참고문헌

[1] V. Vapnik, " The Nature of Statistical Learning Theory," Springer-Verlag, 1995.

[2] William W. Cohen, " Learning Rules that Classify E-Mail," AAAI Spring Symposium on Machine Learning in Information Access, pp.18-25, 1996.

[3] P.Pantel and D.Lin, " SpamCop: A Spam Classification and Organization Program," In Learning for Text Categorization: Papers from the 1998 WorkShop, 1998.

[4] T. Joachims, " Text categorization with support vector machine: Learning with many relvant features," In European Conf. Machine Learning, 1998.

[5] C. Cortes and V. Vapnik, " Support Vector Networks, Machine Learning," Vol.20, pp.273-297, 1995.

[6] Sahami, M., Dumais, S., Heckerman, D. and Horvitz. E, " A Bayesian approach to filtering junk e-mail," AAAI 98Workshop on Text Categorization, 1998.

[7] Salton. G and McGill. M, " Introduction to Modern Information Retrieval," McGraw Hill, 1983.

[8] 조한철, 조근식, " 나이브 베이지안 분류자와 매시지 규칙을 이용한 스팸메일 필터링 시스템," 제29회 춘계학술대회, 한국정보과학회, pp.223-225, 2002.

[9] T. Joachims, SVMLight, http://ais.gmd.de/~thorsten/svm_light, 1998.