카이 제곱 통계량과 지지벡터기계를 이용한 자동 스팸 메일 분류기

이성욱*

*충주대학교

An Automatic Spam e-mail Filter System Using χ^2 Statistics and Support Vector Machines

Songwook Lee*

*Chungju National University

E-mail : leesw@cjnu.ac.kr

요 약

우리는 지지벡터기계를 이용하여 스팸 이메일을 자동으로 분류하는 시스템을 제안한다. 단어의 어휘 정보와 품사 태그 정보를 지지벡터기계의 자질로 사용한다. 우리는 카이 제곱 통계량을 이용하여 유용한 자질을 선택한 후 각각의 자질을 문서 빈도(TF)와 역문헌빈도(IDF) 값으로 표현하였다. 자질들을 이용하여 SVM을 학습한 후, SVM 분류기는 각각의 이메일의 스팸 유무를 결정한다. 실험 결과, 웹메일 시스템에서 수집한 이메일 데이터에 대해 약 82.7%의 정확률을 얻었다.

ABSTRACT

We propose an automatic spam mail classifier for e-mail data using Support Vector Machines (SVM). We use a lexical form of a word and its part of speech (POS) tags as features. We select useful features with χ^2 statistics and represent each feature using text frequency (TF) and inversed document frequency (IDF) values for each feature. After training SVM with the features, SVM classifies each email as spam mail or not. In experiment, we acquired 82.7% of accuracy with e-mail data collected from a web mail system

키워드

spam mail filter, support vector machine, chi square statistics

Ⅰ. 서론

인터넷의 발달과 웹 메일 서비스의 보급으로 인해 전자우편은 사용하기 쉽고, 빠른 편리함으로 널리 사용되고 있다. 그러나 인터넷의 상업적 이용과 개인정보를 이용한 범죄의 목적으로 상당량의 스팸메일이 홍수를 이루고 있다.

스팸메일이란 불특정 다수에게 수신자의 동의 없이 발송되며, 수신자에게 불필요한 정보를 담고 있는 전자우편을 뜻한다. 이러한 스팸메일은 사용자의 불편을 초래할 뿐만 아니라 이메일 시스템에 상당한 부하를 주고 있다. 이러한 스팸메일을 차단하는 스팸메일 필터링에 관한 연구가 활발히 진행되고 있는데, 대부분의 연구는 베이지안 분류기를 기반으로 하고 있으며1-5, 그 외, 마코프 랜덤 필드 (Markov Random Field) 모델6과 k-Nearest Neighbor(k-NN) 방법[7]을 이용한 연구가 있다.

가중치가 부여된 베이지안 분류기[1]는 메일 분류를 위한 전처리 단계와 사용자의 행동을 반영할 수 있는 지능형 에이전트가 결합된 형태의 시스템을 제안하였다. 자질들의 독립을 가정하는 나이브 베이지안 분류기를 확장한 Less Naive Bayes(LNB) 방법과 메일 발송 서버 주소를 이용하여 메일을 분류하는 SMTP 경로 분석 분류기의 통합을 제안한 방법도 있었다[2]. 이러한 독립적

분류기의 통합은 다양한 자질의 조합으로 분류기의 정확도를 향상시킬 수 있는 장점이 있다.

문자열 기반 베이지안 분류기[3]는 각 클래스별로 문자열의 확률을 추정하는 모델을 생성하고 이를 분류기로 이용하였다.

필터 시스템의 정확률과 오류율을 손실 비율에 따라 다른 가중치를 적용하여 계산한다[4]. 정확률에서는 정상 메일로 분류한 것에 가중치를 부여하고 오류율에서는 정상메일을 스팸메일로 분류한 경우에 가중치를 부여하여 정상메일이 스팸메일로 분류될 때의 오류를 스팸메일이 정상메일로 분류될 때의 오류보다 큰 오류로 보았다.

다이그래믹(digramic) 베이시안 분류기를 이용한 시스템5은 각 클래스에서 최대 엔트로피 원리를 이용한 파라미터를 계산하여 그 값을 베이지안 기법에 이용하여 문서의 클래스를 결정한다.

마코프 랜덤 필드 모델을 이용한 스팸 메일 필터 시스템[6]은 윈도우 사이즈를 5로 하는 직교 스파스 바이그램(Orthogonal Sparse Bigram) 자질을 이용하였는데, 인접한 5개의 단어를 두 단어씩 묶어 자질로 이용하였다.

k-NN 분류기[7]는 거리에 따른 가중치와 정확도에 따른 가중치를 적용하였는데 가중치가 적용된 유클리디안 거리 함수를 학습 문서와 테스트 문서 사이의 유사도 측정에 사용하였고 새 문서를 분류할 때, 이전 학습 문서들 중 정확한 분류에 기여한 학습 문서의 가중치를 높여줌으로써 좋은 자질에 가중치를 주었다.

스팸 메일 분류의 경우, 스팸인 메일과 정상인 메일을 구분하는 이진 분류의 성격을 가지고 있으므로 본 연구에서는 이진 분류기 중에서 가장 성능이 좋다고 알려진 지지 벡터 기계(Support Vector Machine)를 스팸 메일 필터 시스템에 이용하는 것을 제안한다.

기계 학습에서 적절한 자질의 선택은 시스템의 성능에 많은 영향을 끼친다. 그러나 본 연구에서는 지지벡터 기계 자체가 가진 자질 선택성을 이용하여 추출된 어휘와 품사 자질만 가지고 실험을 한다. 다음 그림 1은 제안하는 시스템의 구조도이다.

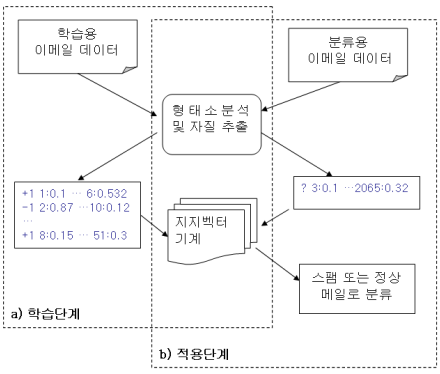

그림 1. 제안 시스템 구조도.

제안하는 스팸메일 필터 시스템은 크게 두 단계로 나뉜다. 먼저 학습단계에서는 학습용 이메일 데이터로부터 지지벡터 기계의 학습에 사용할 수 있는 자질을 추출하여야 한다. 학습용 이메일 데이터는 형태소 분석 단계를 거쳐 어휘/품사 쌍으로 자질을 이룬다. 각 자질은 해당하는 차원의 축을 이루며 각 자질의 가중치가 그 차원의 값이 된다. 벡터로 이루어진 데이터가 만들어지면 지지벡터 기계를 학습한다.

지지벡터 기계가 학습되고 나면 분류용 이메일 데이터를 스팸인지 아니면 정상 메일인지 분류할 수 있게 된다. 학습시와 마찬가지로 분류용 이메일 데이터는 형태소 분석 단계와 자질 추출 단계를 거쳐 다차원 상의 한 점을 이루는 벡터 데이터가 되고 이를 지지벡터 기계가 스팸 또는 정상 메일로 분류하게 된다.

본 논문의 구성은 다음과 같다. 2장에서는 지지벡터 기계의 학습에 사용된 자질을 설명하며, 3장에서는 지지벡터 기계에 대해 소개하며 어떻게 제안된 스팸메일 필터 시스템에 적용하였는지 설명한다. 4장에서는 실험을 통해 제안된 방법의 성능을 보이고 5장에서 결론을 내린다.

II. 자질과 카이 제곱 통계량

본 연구에서는 하나의 웹 메일 시스템의 계정에서 수집한 EML형식의 837개의 스팸메일과 600개의 정상적인 메일을 실험에 이용한다. 수집된 메일은 한국어형태소 분석기를 이용하여 자동으로 품사를 부착하였다. 다음 그림 2는 품사 부착 전의 메일과 품사 부착 후의 메일 데이터의 예를 나타낸다.

[목표미달성시수강료50%환급] 3개월 후
당신의 영어회화 실력을 보장합니다!

a) 품사 부착 전

목표/NNG+미달성/NNG+시/XSN+수강료/NNG
3/SN+개월/NNB
후/NNB
당신/NP + 의/J
영어회화/NNG
실력/NNG + 을/J
보장/NNG + 하/XSV + ㅂ니다/E

b) 품사 부착 후

그림 2. 품사 부착 전후의 이메일 예.

우리는 수집된 이메일 파일을 마이크로소프트사의 WordBreaker2007을 이용하여 자동으로 품사를 부착하였으며, 품사가 부착된 어휘/품사 쌍을 자질로 사용하였다. 따라서 가능한 자질의 종류는 수집된 이메일에서 발견되는 모든 어휘/품사 쌍이 되며, 매우 많은 수의 자질이 나타나게 된다.

이러한 자질들 중에서는 스팸 메일을 결정하는 데 기여를 하는 자질이 있기도 하지만 그렇지 않은 경우나 오히려 방해가 되는 자질들도 존재를 하게 된다. 따라서 우리는 카이 제곱 통계량을 이용해서 자질을 선택한다. 카이 제곱 통계량을 계산하는 식은 다음과 같다 [11].

$$\chi^2(f,s)=\frac{(A+B+C+D)\times(AD-BC)^2}{(A+B)\times(A+C)\times(B+D)\times(C+D)} \quad (1)$$

A는 스팸메일 s 중에 자질 f를 포함하고 있는 문서의 수이고, B는 범주 s 이외의 문서, 즉 정상 메일 중 속해 있는 문서 중에 자질 f를 포함하고 있는 문서의 수이다. 또한, C는 스팸메일 s에 속해 있는 문서 중에 자질 f를 포함하지 않는 문서의 수이며, D는 범주 s외의 문서 중에 자질 f를 가지고 있지 않는 문서의 수이다. 자질 f와 범주 s가 완전히 독립적이면 0의 값을 갖는다. 하나의 자질에 대해 카이 제곱 통계량의 값을 결정하는 방법은 전체 범주에 대한 평균값을 사용하는 방법과 전체 범주에 대해 최대값을 사용하는 방법이 있을 수 있다. 우리는 이것을 이진 분류에 사용하므로 각 자질 당 하나의 값만 사용한다.

각각의 자질에 가중치를 부여하는 방법은 이진 가중치, 용어 및 역문헌 빈도 (Term Frequency-Inverse Document Frequency) 가중치, 용어 및 역범주 빈도 (Term Frequency- Inverse Category Frequency) 등 여러가지가 있으면 본 연구에서는 일반적으로 가장 좋은 성능을 보이는 TF-IDF 가중치 방법을 사용한다.

스팸 메일 필터기에 적용하기 위해 TF-IDF 값을 계산하는 경우, 용어(term)는 자질로, 문서(document)는 이메일로 범주(category)는 스팸메일과 정상적 메일로 간주하여 계산한다.

IV. 지지벡터 기계

지지벡터기계(Support Vector Machine)는 두 개의 범주를 구분하는 문제를 해결하기 위해 1995년에 Vapnik8에 의해 소개된 학습기법으로 그림 3과 같이 선형공간(hyper-space)에서 두 개의 클래스의 구성 데이터들을 가장 잘 분리해 낼 수 있는 결정면(decision surface)을 찾는 모델이다.

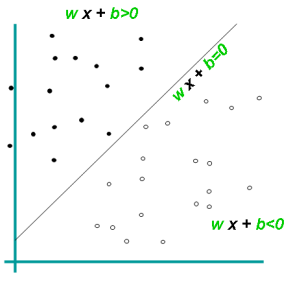

그림 3. 선형 공간에서의 결정면.

선형 분리가 가능한 공간에서의 결정면은 초월면(hyper-plane) H : y=w·x + b = 0 이며 이 초월면에 평행하고 동일 거리에 있는 두 개의 초월면은 아래 식의 H1,H2와 같으며, H1와 H2사이에 어떠한 데이터 포인트도 존재하지 않는 조건을 만족시키며 H1와 H2사이의 거리는 최대가 된다.

H1 : y=w·x + b = +1,
H2 : y=w·x + b = －1.

H1와 H2사이의 거리를 최대로 만드는 것이 지지벡터 기계의 학습 목적이 된다. 따라서 H1에는 양의 값을 갖는 데이터가 존재하게 되고 H2에는 음의 값을 갖는 데이터가 존재하게 되는데, 이러한 데이터들을 지지벡터(support vectors)라 부르며 이들이 분리 경계면을 결정하는 역할을 한다. 다른 데이터들은 H1와 H2를 교차하지 않도록 분리 경계면 주위로 이동되거나 제거된다. H1와 H2사이의 거리 M을 최대로 하기 위해서 H1와 H2사이에 어떠한 데이터 포인트도 존재하지 않도록 하면서 ||w||을 최소화시키면 된다.

w·x + b ≥ +1 for yi=+1,
w·x + b ≤ －1 for yi= －1.

지지벡터기계의 문제는 이러한 w 와 b를 찾아내는 문제이며, 이것은 2차 프로그래밍(quadratic programming) 기술에 의해 풀 수 있다[8].

문서 분류 분야에서 좋은 성능을 보여 주고 있는 지지벡터기계를 우리는 스팸 메일 분류에 사용하였다. 지지벡터기계는 이진 분류기이므로 우리는 스팸메일과 정상메일을 분류하기 위해 하나의 모델만 학습하면 된다. 스팸 메일인 경우 양(+1)의 자질을 정상적 메일 경우 음(-1)의 자질을 부여하였다. 지지벡터기계의 학습을 위한 자질은 2장에서 설명한 어휘/품사 쌍의 자질들의 가중치로 벡터를 구성하였다. 본 연구에서는 LIBSVM[9]을 이용하였고 여러 가지 커널에 대해 반복 실험한 결과 커널의 영향을 거의 받지 않으므로 선형 커널을 이용하여 학습하였다.

V. 실험 및 결과

본 연구에서 사용한 이메일 데이타는 웹 메일 시스템으로부터 얻은 EML형식의 837개의 스팸메일과 600개의 정상적인 이메일을 실험에 이용하였다. 1,317개의 데이터를 학습 데이터로, 120개의 데이터를 평가 데이터로 사용한다. 스팸 메일 필터 시스템은 주어진 이메일이 스팸인지 아닌지를 판별하는 시스템이다. 따라서 제안하는 시스템은 각각의 이메일 데이터를 얼마나 제대로 분류

했는지를 나타내는 척도로 정확률(accuracy)을 사용하며, 실험 결과 88.7%의 정확률을 얻었다.

표1은 제안 시스템과 스팸성 자질과 URL 자질의 공동 학습을 통해 최대 엔트로피 학습 방법으로 스팸 메일을 판별하는 시스템[10]과 비교한 것이다. 카이 제곱 통계량을 이용하여 자질을 선택했을 때, 자질 선택 전보다 약 2.2%의 성능 향상을 보인다. 수치상으로 제안하는 시스템이 좋은 성능을 보이지만, 사용하는 자질의 종류와 실험에 사용한 데이터의 양이 서로 달라 정확한 비교는 어렵다. 비교 시스템이 사용하는 자질의 다양성과 최대 엔트로피 모델의 계산량과 비교할 때, 어휘/품사 쌍 자질만 사용하는 제안 시스템의 성능이 비교 시스템의 성능보다 나은 것을 알 수 있다.

표 1. 다른 시스템과의 정확률 비교

	정확률(%)	비고
제안시스템	80.5	모든 자질
	78.4	$\chi^2>3$
	82.7	$\chi^2>4$
	77.8	$\chi^2>6$
비교시스템[10]	79.6	

대부분의 오류는 형태소 분석기의 오류와 불충분한 데이터로 인해 발생하고, 수집된 데이터의 순결성이 떨어지는 데서 발생했다. 또한 이메일의 내용이 멀티미디어 컨텐츠만 포함하고 있는 경우, 멀티미디어 내용을 알 수 없으므로 스팸 메일 분류를 어렵게 한다. 가장 어려운 오류 유형은 홈쇼핑 사이트 등에서 발송하는 쇼핑 정보의 경우 일반적인 스팸 메일의 특성을 많이 가지고 있음에도 불구하고, 사용자의 선택에 따라 정상적인 메일로 분류되므로 오류로 발생한다. 이와 같은 오류는 사용자의 사용행태를 반영할 수 있는 매커니즘에 대한 연구가 필요하다.

VI. 결론 및 향후과제

본 논문에서는 범람하는 스팸 메일을 차단하기 위해, 어휘/품사 쌍의 자질을 이용하고 카이제곱 통계량을 이용하여 자질을 선택한 후, 선택된 자질로 지지벡터기계를 학습하여 자동으로 스팸메일을 걸러낼 수 있는 스팸 메일 필터 시스템을 제안하였다. 실험에 사용된 이메일은 웹 메일 시스템에서 자동으로 수집하였으며, 실험 결과 82.7%의 정확률을 얻었다. 대부분의 오류는 형태소 분석기의 오류에서 발생하며, 멀티미디어 데이터를 포함한 이메일의 경우, 일일이 인코딩된 데이터를 디코딩할 수 없으므로 스팸 메일 분류에 어려움이 있다. 모든 어휘/품사 자질을 사용할 경우 너무 많은 자질이 사용되므로 URL정보 등과 같이 좀 더 유용한 자질을 추가 및 선택하는 방법에 대한 연구가 필요하다. 본 실험에서는 웹 메일 시스템으로부터 그리 많지 않은 양의 데이터로 실험하였으므로 자료 부족 문제도 발생하였다. 따라서 더 많은 데이터의 수집이 필요하며 제안된 시스템과 인터넷 이메일 에이전트를 결합하여 실생활에 유용한 이메일 사용 환경을 구현할 필요가 있다.

참고문헌

[1] Keselj, V., Milios, E., Tuttle, A., Wang, S., Zhang, R. "TREC 2005 Spam Track: Spam Filtering Using N-gram-based Techniques", Proceedings of Text REtrieval Conference, 2005.

[2] 김현준, 정재은, 조근식, “가중치가 부여된 베이지안 분류자를 이용한 스팸 메일 필터링 시스템”, 정보과학회논문지, 31권 8호, 2004, pp.1092-1100.

[3] Segal, R. "IBM SpamGuru on the TREC 2005 Spam Track", Proceedings of Text REtrieval Conference, 2005.

[4] Brakto, Al, Filipic, B., "Spam Filtering Using Character-Level Markov Models: Experiments for the TREC 2005 Spam Track", Proceedings of Text REtrieval Conference, 2005.

[5] Breyer, L. A. "DBACL at the TREC 2005", Proceedings of Text REtrieval Conference, 2005.

[6] Assis, F., Yerazunis, W., Siefkes, C., Chhabra, S., "CRM114 versus Mr. X: CRM114 Notes for the TREC 2005 Spam Track", Proceedings of Text REtrieval Conference, 2005.

[7] Cao, W., An, A., Huang, X. "York University at TREC 2005: SPAM Track", Proceedings of Text REtrieval Conference, 2005.

[8] V. Vapnik. The nature of statistical learning theory, Springer, NewYork, 1995.

[9] http://www.csie.ntu.edu.tw/~cjlin/libsvm

[10] 공미경, 이경순, “스팸성 자질과 URL 자질의 공동 학습을 이용한 최대 엔트로피 기반 스팸 메일 필터 시스템”, 정보처리학회논문지B, 15-B권 1호, pp61-68, 2008.

[11] Yang, Yiming and Jan O. Pedersen. A comparative study on Feature selection in text categorization. *In proceedings of the 14th International conference on Machine Learning,* 1997.