DOI: 10.3745/KIPSTB.2010.17B.3.249

카이제곱 통계량과 지지벡터기계를 이용한 스팸메일 필터

이 성 욱†

요 약

본 논문은 지지벡터기계를 이용하여 스팸메일을 자동으로 분류하는 시스템을 제안한다. 이메일에 포함된 단어의 어휘 정보와 품사 태그 정보를 지지벡터기계의 자질로 사용한다. 우리는 카이제곱 통계량을 이용하여 자질을 선택한 후 각각의 자질을 TF, TF-IDF, 이진 가중치 등으로 표현하여 실험하였다. 카이제곱 통계량을 이용하여 선택된 자질들을 이용하여 SVM을 학습한 후, SVM분류기는 각각의 이메일의 스팸 여부를 결정한다. 실험 결과, 선택되어진 자질들이 성능향상을 가져왔으며, TREC05-p1 스팸 말뭉치에 대해 약 98.9%의 정확도를 얻었다.

키워드 : 스팸메일 분류기, 지지벡터기계, 카이제곱 통계량, 자질 선택

Spam Filter by Using χ^2 Statistics and Support Vector Machines

Songwook Lee†

ABSTRACT

We propose an automatic spam filter 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 and select features by chi square statistics. We represent each feature by TF(text frequency), TF-IDF, and binary weight for experiments. After training SVM with the selected features, SVM classifies each e-mail as spam or not. In experiment, the selected features improve the performance of our system and we acquired overall 98.9% of accuracy with TREC05-p1 spam corpus.

Keywords : Spam Mail Filter, Support Vector Machine, Chi Square Statistics, Feature Selection

1. 서 론

인터넷의 발달과 웹 메일 서비스의 보급으로 인해 전자우편은 그 편리함으로 인해 실생활에 널리 사용되고 있다. 그러나 인터넷의 상업적 이용과 개인정보를 이용한 범죄의 목적 등으로 매일 수신되는 스팸메일이 점점 많아지고 있다.

스팸메일이란 불특정 다수에게 수신자의 동의 없이 발송되며, 수신자에게 불필요한 정보를 담고있는 전자우편을 뜻하며, 이러한 스팸메일은 사용자의 불편을 초래할 뿐만 아니라 이메일 시스템에 상당한 부하를 준다. 이러한 스팸메일을 차단하는 스팸메일 필터링에 관한 연구가 활발히 진행되고 있는데, 대부분의 연구는 베이지안 분류기를 기반으로 하고 있으며[1-5], 그 외, 마코프 랜덤 필드(Markov Random Field) 모델[6]과 k-Nearest Neighbor(k-NN) 방법[7], 최대 엔트로피 모형을 이용한 방법[10], 지지벡터기계(Support Vector Machine)를 이용한 연구[12, 13]가 있다.

가중치가 부여된 베이지안 분류기[2]는 정보통신부의 개정안을 준수하는 메일 분류를 위한 전처리 단계와 사용자의 행동을 학습하는 지능형 에이전트가 결합된 형태의 시스템을 제안하였고 나이브 베이지안 분류기보다 재현율에서 우수함을 보였다. 자질들의 독립을 가정하는 나이브 베이지안 분류기를 확장한 Less Naïve Bayes(LNB) 방법과 메일 발송 서버 주소를 이용하여 메일을 분류하는 SMTP 경로 분석 분류기의 통합을 제안한 방법[3]도 있는데, 이러한 독립적 분류기의 통합은 다양한 자질의 조합으로 분류기의 정확도를 향상시킬 수 있는 장점이 있다. 문자열 기반 베이지안 분류기[4]는 일반적으로 사용되는 단어에 기반한 자질 대신에 각 클래스별로 문자열의 확률을 추정하는 모델을 생성하고 이를 분류기로 이용하였는데, 문서의 클래스는 테스트 문서가 각 클래스에 포함될 조건확률을 계산하여 결정한다. 다이그래믹(digramic) 베이시안 분류기를 이용한 시스템[5]은 각 클래스에서 최대 엔트로피를 이용한 파라미터를 계산하여 그 값을 베이지안 분류기법에 이용하여 문서의 클래스

※ 이 논문은 충주대학교 대학구조개혁지원사업비(교육과학기술부 지원)의 지원을 받아 수행한 연구임.
† 정 회 원 : 국립충주대학교 컴퓨터정보공학과 조교수
논문접수 : 2009년 12월 18일
수 정 일 : 1차 2010년 3월 5일, 2차 2010년 3월 19일
심사완료 : 2010년 3월 22일

를 결정한다. 이러한 베이지안 분류기에 기반한 시스템은 베이지안 분류기가 각 자질의 독립을 가정하고 있으므로 새로 들어오는 문서에 의해 각 자질의 가중치만 새로 계산하면 되며, 전체 학습 데이터를 다시 학습할 필요가 없는 장점이 있다. 반면 각 자질의 독립을 가정하고 있는 분류기의 특성상 문맥 정보를 반영할 수 없는 단점이 있으며, 보통 이를 극복하기 위해 다양한 분류기의 결합을 시도한다.

마코프 랜덤 필드 모델을 이용한 스팸메일 필터 시스템 [6]은 CRM114[1] 분류기와 윈도우 사이즈를 5로 하는 직교 스파스 바이그램(Orthogonal Sparse Bigram) 자질을 이용하여 마코프 랜덤 필드 모델이 스팸메일 필터와 같은 분류기로 좋은 성능을 보인다는 것을 보였는데, 직교 스파스 바이그램이란 인접한 5개의 단어열 중 첫 번째 단어와 나머지 네 개의 단어를 각각 바이그램으로 묶어 자질로 이용하는 것을 말한다.

k-NN 분류기를 이용한 [7]은 거리 가중치와 정확도 가중치를 학습하여 스팸메일 분류에 사용하였다. 정확도 가중치는 학습문서가 스팸 특성의 변화에 더 잘 적응하도록 학습문서 유지관리에도 사용되었는데, 새 문서를 분류할 때, 이전 학습문서들 중 정확한 분류에 기여한 학습문서의 가중치를 높여줌으로써 좋은 자질에 가중치를 주었다. 가중치가 적용된 거리 함수를 학습문서와 테스트 문서 사이의 유사도 측정에 사용하여 스팸메일을 분류하였다.

[10]은 스팸메일의 제목과 본문에 나타나는 특정한 패턴을 스팸성 자질로 정의하고 URL 자질과 공동학습을 통해 두 분류기를 결합하였으며, 각 자질의 확률값은 최대 엔트로피 모형을 이용하여 계산하였으나 다른 시스템과 비교하여 좋은 성능을 보여주지 못했다.

스팸메일 분류의 경우, 스팸메일과 정상메일을 구분하는 이진 분류의 성격을 가지고 있고, 지지벡터기계(SVM)를 이용한 방법이 많이 연구되었다[12,13]. 실용적인 관점에서 지지벡터기계는 온라인에서 사용하기 어려운데, 그 이유는 새로운 데이터가 들어오면 전체 모델을 다시 학습해야하기 때문이다. 이러한 단점을 해결하기 위해 [12]는 SVM의 성능을 유지하면서도 온라인상에서 사용할 수 있도록 계산량을 감소시킨 방법을 제안하였고 좋은 결과를 얻었다.

필터 시스템의 성능평가와 관련해서는 정확률과 오류율을 손실 비율에 따라 다른 가중치를 적용하여 계산한다[4,15]. 정확률에서는 정상메일로 분류한 것에 가중치를 부여하고 오류율에서는 정상메일을 스팸메일로 분류한 경우에 가중치를 부여하여 정상메일이 스팸메일로 분류될 때의 오류를 스팸메일이 정상메일로 분류될 때의 오류보다 큰 오류로 보았다. 대표적인 평가방법은 TREC05부터 사용된 HM, SM, LM 방법[15]과 TREC07부터 사용되고 있는 1-ROCA 방법[15, 16]이 있다.

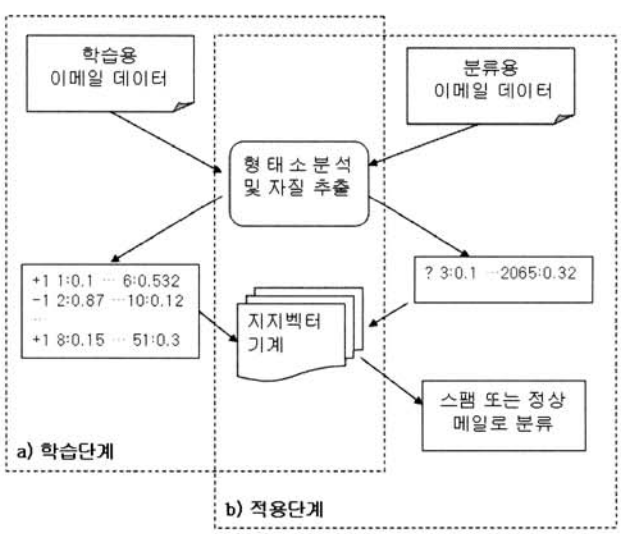
본 연구에서는 지지벡터기계에 사용하는 자질을 카이제곱 통계량을 이용하여 선택하는 방법[14]을 스팸메일 필터 시스템에 이용하는 것을 제안한다. 스팸메일 필터 시스템은 수신된 이메일을 자동으로 스팸메일과 정상메일로 분류하는 이진 분류 시스템이다.

기계 학습에서 적절한 자질의 선택은 시스템의 성능에 많은 영향을 끼친다. 스팸메일 필터 시스템에서는 URL자질과 n-gram 자질을 주로 사용하는데, 본 연구에서는 어휘/품사 자질만 사용하며 각 자질의 가중치로 TF, TF-IDF, 이진 가중치 등을 사용하여 실험을 한다. 다음 (그림 1)은 제안하는 시스템의 구조도이다.

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 정상메일인지 분류할 수 있게 된다. 학습 때와 마찬가지로 분류용 이메일 데이터는 형태소 분석 단계와 자질 추출 단계를 거쳐 다차원 상의 한 점을 이루는 벡터 데이터가 되고 이를 지지벡터 기계가 스팸 또는 정상메일로 분류하게 된다.

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



(그림 1) 제안 시스템 구조도

2. 자질과 카이제곱 통계량

본 연구에서는 TREC05-p1 데이터[19]를 실험에 이용한다. 수집된 메일은 형태소 분석기를 이용하여 자동으로 품사를 부착하였다. 다음 (그림 2)는 품사 부착 전의 메일과

[1] http://crm114.sourceforge.net

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

a) 품사 부착 전

목표미달성시수강료/NN + 50%/NU
환급/NN
3/NU+개월/NN+후/SF
당신/NN + 의/JO
영어회화/NN
실력/NN + 을/JO
보장/NN + 하/SF + ㅂ니다/EM

b) 품사 부착 후

(그림 2) 품사 부착 전후의 이메일 예

품사 부착 후의 메일 데이터의 예를 나타낸다.

본 연구에서는 수집된 이메일 파일을 HAM[17]과 Montylingua[18]을 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 스팸메일을 결정하는 데 기여를 하는 자질이 있기도 하지만 그렇지 않은 경우나 오히려 방해가 되는 자질들도 존재를 하게 된다. 불필요한 자질을 제거하기 위해 카이제곱 통계량을 이용해서 자질을 선택한다. 카이제곱 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 정상메일 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 카이제곱 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 여러 가지가 있다. 본 연구에서는 TF 가중치, TF-IDF 가중치, 이진 가중치를 각각 사용하여 실험한다. 스팸메일 필터기에 적용하기 위해 TF-IDF 값을 계산하는 경우, 용어(term)는 자질로, 문서(document)는 이메일로, 범주(category)는 스팸메일과 정상메일로 간주하여 계산한다.

3. 지지벡터 기계

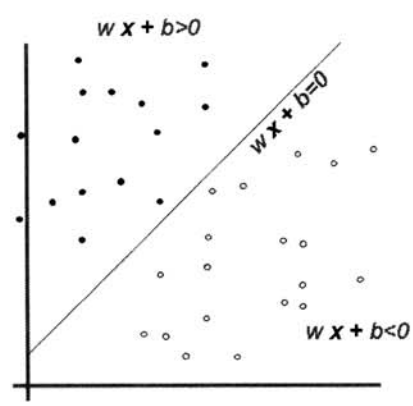
지지벡터 기계(Support Vector Machine)는 두 개의 범주를 구분하는 문제를 해결하기 위해 1995년에 Vapnik[8]에 의해 소개된 학습기법으로 (그림 3)과 같이 초월공간(hyper-space)에서 두 개의 클래스의 구성 데이터들을 가장 잘 분리해 낼 수 있는 결정면(decision surface)을 찾는 모델이다.

(그림 4)와 같이, 선형 분리가 가능한 공간에서의 결정면은 초월면(hyper-plane) H : $y=w\cdot x + b = 0$ 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.

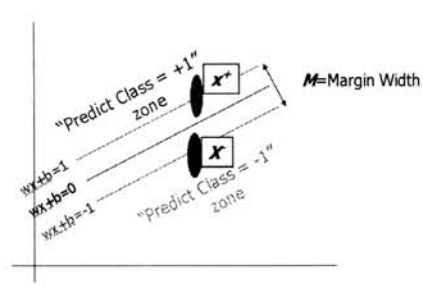
$$H1 : y=w\cdot x + b = +1,$$

$$H2 : y=w\cdot x + b = -1.$$

H1와 H2사이의 거리를 최대로 만드는 것이 지지벡터 기계의 학습 목적이 된다. 따라서 H1에는 양의 값을 갖는 데



(그림 3) 선형 공간에서의 결정면[21]



(그림 4) 최대 마진폭을 가지는 H1과 H2[21]

이터가 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.

$$w \cdot x + b \geq +1 \text{ for } yi=+1,$$

$$w \cdot x + b \leq -1 \text{ for } yi=-1.$$

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

문서 분류 분야에서 좋은 성능을 보여 주고 있는 분류기인 지지벡터기계를 우리는 스팸메일 분류에 사용하였다. 지지벡터기계는 이진 분류기이므로 우리는 스팸메일과 정상메일을 분류하기 위해 하나의 모델만 학습하면 된다. 스팸메일인 경우 양(+1)의 자질을, 정상메일인 경우 음(-1)의 자질을 부여하였다. 지지벡터 기계의 학습을 위한 자질은 2장에서 설명한 어휘/품사 쌍의 자질들이 벡터의 각 차원을 구성하며 TF, TF-IDF, 또는 이진 가중치 등이 각 차원의 값이 되어 벡터를 구성한다. 본 연구에서는 LIBSVM[9]을 이용하였고 여러 가지 커널에 대해 반복 실험 결과 비교적 좋은 성능을 보인 선형 커널을 이용하여 학습하였다.

4. 실험 및 결과

본 연구에서 TREC05-p1 말뭉치[19]의 일부분을 사용하였다. TREC05-p1 말뭉치는 92,189개의 이메일로 구성되어 있으며 정상메일과 스팸메일의 비율을 조정한 5개의 부분집합이 존재한다. 기계학습의 편이를 위하여 그 중에서 랜덤 추출한 6,572개의 데이터를 학습데이터로 사용한다. 평가데이터로는 TREC05-p1의 5개의 부분집합을 학습데이터와 겹치지 않도록 각각 1/100의 확률로 랜덤 추출한 데이터 3,168개를 사용한다. 다음 <표 1>은 실험에 사용된 말뭉치 구성을 보인다.

스팸메일 필터 시스템은 주어진 이메일이 스팸인지 아닌지를 판별하는 시스템이다. 우리는 제안 시스템의 평가 척도로 일반적인 시스템에서 많이 사용하는 정확률, 재현율, 정확도(accuracy) 등과 스팸메일 분류기에 사용하는 정상메일오류율-Hm(Ham misclassification rate), 스팸메일오류율-Sm(Spam misclassification rate), (1-ROCA)% 방법을 사용한다. [16]에서 소개된 1-ROCA 방법은 ROC(Receiver Operating Characteristic) 곡선 아래의 면적을 이용하여 계산하는데, 그 값은 어떤 스팸메일 분류기(classifier)가 하나의 메시지를 정상메일이 아니라 스팸메일이라고 판단하였을 때 오류가 발생할 확률값이며, '0%'의 값을 가질 때 완벽한 시스템이 된다. 최근 스팸메일 분류 시스템의 성능을 나타내기 위해 많이 사용되고 있는 평가방법이다.

$$Hm(\%) = \frac{\text{정상메일을 스팸메일로 분류한 개수}}{\text{정상메일의 수}} * 100$$

$$Sm(\%) = \frac{\text{스팸메일을 정상메일로 분류한 개수}}{\text{스팸메일의 수}} * 100$$

$$(1\text{-}ROCA)\% = 1 - \text{Area under ROC curve}$$

먼저 최적의 자질 개수를 선택하기 위해 카이제곱 통계량에 따른 성능을 살펴보자. 다음 <표 2>는 카이제곱 통계량에 따른 정확도를 나타내며 자질 벡터의 가중치로 TF-IDF를 사용하고 ham50 평가 데이터로 평가한 결과이다.

<표 2>의 결과와 같이 전체 약 25만 6천여 개의 자질들 중 카이제곱 통계량을 이용하여 $\chi^2 > 4.5$의 약 4만 5천여 개의 자질을 선택하였을 때 가장 좋은 성능을 보였다.

다음 <표 3>은 자질의 가중치의 종류에 따른 성능을 나타낸다. 자질의 개수는 <표 2>의 결과에 따라 $\chi^2 > 4.5$의 값으로 제한하였다.

<표 3>에서와 같이 스팸메일 분류에 SVM을 사용할 때에는 자질의 가중치를 TF나 TF-IDF보다 이진 가중치를 사용하였을 때 가장 좋은 결과를 보였다.

다음 <표 4>는 카이제곱 통계량을 이용하여 선택한 자질

〈표 1〉 말뭉치 구성

	스팸메일	정상메일	평가데이터 집합(개수)
TREC05-p1 추출	4,821	1,751	Full(813) ham25(568) ham50(701) spam25(493) spam50(603)

〈표 2〉 카이제곱 통계량에 따른 시스템 정확도

평가데이터	정확도(%)	χ^2	자질의 개수
ham50	87.87	$\chi^2 >= 0$	256,225
	88.59	$\chi^2 > 4.5$	45,577
	88.45	$\chi^2 > 5$	44,367

〈표 3〉 자질 가중치의 종류에 따른 성능(701개의 ham50 평가 데이터 사용)

(단위 : %)

가중치	정확률	재현율	F1	Hm	Sm	정확도	(1-ROCA)
TF-IDF	86.90	99.21	93.06	39.39	0.79	88.59	0.2324
TF	99.21	99.21	99.21	2.07	0.79	98.86	0.0122
이진	99.41	99.80	99.61	1.87	0.19	99.43	0.0049

〈표 4〉 이진 가중치를 이용한 시스템 성능(%)

평가데이터	정확률	재현율	F1	Hm	Sm	정확도	(1-ROCA)	X^2
Trec05(full)	98.42	99.32	98.87	1.87	0.68	98.77	0.0013	$X^2 > 4.5$
Trec05(ham50)	99.41	99.80	99.61	1.87	0.19	99.43	0.0049	
Trec05(ham25)	100.0	99.79	99.90	0.0	0.21	99.82	0.0012	
Trec05(spam50)	95.24	99.10	97.17	2.89	0.90	97.84	0.0128	
Trec05(spam25)	94.07	100.0	97.04	1.83	0.0	98.58	0.0046	
Overall	98.42	99.60	99.01	1.98	0.40	98.90	0.0142	$X^2 > 4.5$

을 이진 가중치로 나타내어 자질 벡터를 구성하였을 때, 시스템 성능을 나타낸다.

<표 5>와 <표 6>은 제안 시스템과 관련 연구들을 비교한 것이다. 제안 시스템에서 사용한 학습데이터 개수(6,572개)와 비슷한 양의 학습데이터 개수(6,047개)를 사용한 공미경[10]의 시스템은 최대 엔트로피 모형으로 학습된 스팸성 자질 분류기와 URL 자질 분류기의 공동학습을 통해 결합하여 스팸메일을 판별하는 시스템이다. 선택된 어휘/품사 쌍 자질만 사용하는 제안 시스템의 성능이 스팸성 자질과 URL 자질을 이용하는 비교 시스템보다 월등한 것을 알 수 있다.

<표 6>에 나타난 결과는 비교 시스템과 제안시스템은 학습 데이터와 테스트 데이터의 양이 달라 직접적인 비교는 어렵지만 제안 시스템의 성능을 파악하는데 참고가 된다. 제안 시스템을 TREC05WINNERS[15]와 비교할 때, (1-ROCA)척도로 근소하게 낮다고 할 수 있으며, 비교 시스템이 TREC05-p1 전체 데이터를 이용하여 학습한 것을 감안하면 적은 양의 학습데이터를 가지고도 만족할 만한 결과를 얻었다고 볼 수 있다. 단일 분류기로 최고의 성능을 갖는 ONSVM[12]와 (1-ROCA)%를 가지고 비교하면, 단어만 자질로 이용한 결과보다는 제안 시스템의 성능이 근소하게 앞서고 있으며, 4-grams 자질을 이용한 것보다는 성능이 약간 뒤쳐진다고 할 수 있다. 그러나, 비교 시스템은 Spamassassin 데이터 6,034개의 문서를 이용하여 최적의 SVM 파라미터를 튜닝한 후, TREC05-p1 전체 데이터로 SVM모델을 학습하였으므로, 파라미터 튜닝 과정과 4-grams 자질의 복잡성을 고려할 때, 카이제곱 통계량과 어휘/품사 자질만 이용한 제안 시스템의 효율이 더 높다고 할 수 있다. 마지막으로 현재까지 가장 좋은 성능을 보인 것으로 알려진 53-ENSEMBLE[20]은 53개의 스팸 필터를 결합한 시스템으로 단일 분류기인 제안 시스템과 직접적인 비교는 어렵다.

〈표 5〉 TREC05-p1를 사용한 다른 연구와의 비교(1)

비교 시스템	정확률	재현율	F1	Hm	Sm	정확도	비고
제안시스템	98.42	99.60	99.01	1.98	0.40	98.90	6,572
공미경[10]	74.89	78.63	76.7	21.37	19.68	79.60	6,047

〈표 6〉 TREC05-p1를 사용한 다른 연구와의 비교(2)

비교 시스템	1-ROCA
제안시스템	0.014
TREC05WINNERS[15]	0.019
ONSVM:Words[12]	0.015
ONSVM: 4-grams[12]	0.008
53-ENSEMBLE[20]	0.007

일반적으로 스팸메일 분류에서는 정상메일이 스팸으로 분류되는 Hm오류가 스팸메일이 정상메일로 분류되는 Sm오류보다 더 크게 작용한다. 메일 시스템에서 정상메일을 스팸메일로 잘못 분류해 사용자가 못 읽게 되면 사용자에게 큰 손실이 될 수 있기 때문이다. 따라서 Hm이 Sm보다 낮은 값을 가지는 것이 바람직한 시스템이다. 제안 시스템은 Hm이 Sm보다 큰 값을 가지는데 이는 향후 개선이 필요한 결과이며, 학습데이터 중 스팸메일의 데이터의 양이 정상메일보다 약 2.7배 더 많아서 발생한 것으로 보인다. Hm오류를 줄이기 위해 정상메일의 비율을 증가시키고 정상메일의 특성을 잘 반영할 수 있는 새로운 자질을 연구해야한다. 대부분의 오류는 영어와 한국어 형태소 분석기의 오류에서 발생하는데 정형화되지 않은 입력에 대한 토큰 분리가 잘되지 않아 발생했고 데이터의 순도(purity)가 떨어지는 데서 발생했다. 또한 이메일의 내용이 MIME으로 인코딩된 멀티미디어 컨텐츠만 포함하고 있는 경우, 멀티미디어 내용을 알 수 없어 스팸메일 분류를 어렵게 한다. 가장 어려운 오류 유형은 사용자가 가입한 포털 사이트나 홈쇼핑 사이트 등에서 발송하는 안내 메일과 쇼핑정보 메일의 경우 일반적인 스팸메일의 특성을 많이 가지고 있음에도 불구하고, 사용자의 선택에 따라 정상적인 메일로 분류되므로 오류로 작용한다. 이와 같은 오류를 해결하기 위해 사용자의 이메일 수신 행태를 반영할 수 있는 사용자 프로파일(profile)에 대한 연구가 필요하다.

5. 결론 및 향후 과제

본 논문에서는 범람하는 스팸메일을 차단하기 위해, 어휘/품사 쌍의 자질을 이용하여 지지벡터 기계를 학습하여 자동으로 스팸메일을 걸러낼 수 있는 스팸메일 필터 시스템을 제안하였다. 어휘/품사 쌍의 자질은 각 자질의 카이제곱 통계량을 이용하여 선택하였으며 자질을 선택하기 전보다 선택한 후에 시스템의 성능이 향상되었다. 선택된 자질의 가중치로는 이진 가중치가 TF와 TF-IDF 가중치를 사용하였을 때보다 더 나은 결과를 보였다. 실험에 사용된 이메일은 TREC05-p1 데이터에서 추출한 것을 사용하였으며, 실험 결과 TREC05-p1의 평가 데이터에 대해 98.9%의 정확도를 얻었다. 본 실험에서는 TREC05-p1 데이터의 약 1/15의 데이터만 이용하여 학습하였으나 전체 데이터를 이용한 다른 시스템들에 비교할 만한 성능을 얻었다. 대부분의 오류는 형태소 분석기의 오류에서 발생하였고, 멀티미디어 데이터를

포함한 이메일의 경우, MIME 데이터를 해석할 수 없어 MIME 데이터를 학습에서 제외하였으므로 스팸메일 분류에 어려움이 있다. 어휘/품사 자질 이외의 URL 자질과 n-grams 자질 등을 추가하거나 좀 더 유용한 자질을 획득하는 방법에 대한 연구가 필요하며 SVM 학습에 최적의 파라미터 값을 찾기 위한 튜닝 과정도 필요하다. 또한 제안 시스템의 Hm 오류를 낮추기 위한 연구가 필요하다. 추후, 제안하는 방법을 스팸성 블로그를 분류하는데 이용하여 제안 방법을 검증할 필요가 있다. 제안된 시스템과 인터넷 이메일 에이전트를 결합하여 실생활에 유용한 이메일 사용 환경을 제공할 수 있을 것이다.

참 고 문 헌

[1] V. Keselj, E. Milios, A. Tuttle, S. Wang, and R. Zhang. "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 스팸메일 필터링 시스템," 정보과학회논문지, 31권 8호, pp.1092-1100, 2004

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

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

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

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

[7] W. Cao, A. An, and X. Huang. "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 스팸메일 필터 시스템," 정보처리학회논문지B, 15-B권 1호, pp.61-68, 2008.

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

[12] D. Sculley, Gabriel M. Wachman. "Relaxed online SVMs for spam filtering," Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp.415-422, 2007.

[13] H. Drucker, V. Vapnik, and D. Wu. "Support vector machines for spam categorization," IEEE Transactions on Neural Networks, Vol.10, No.5, pp.1048-1054, 1999.

[14] 은종민, 이성욱, 서정연, "지지벡터기계(Support Vector Machines)를 이용한 한국어 화행분석," 정보처리학회논문지, Vol.12-B, No.3, pp.365-368, 2005.

[15] G. V. Cormack and T. R. Lynam. "TREC 2005 spam track overview," The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings, 2005.

[16] G. V. Cormack and T. R. Lynam. "On-line supervised spam filter evaluation," Technical report, David R. Cheriton School of Computer Science, University of Waterloo, Canada, 2006.

[17] http://nlp.kookmin.ac.kr/HAM/kor/index.html

[18] http://web.media.mit.edu/~hugo/montylingua

[19] http://plg.uwaterloo.ca/~gvcormac/treccorpus/

[20] T. Lynam, G. Cormack, and D. Cheriton. "On-line spam filter fusion," Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp.123-130, 2006.

[21] Martin Law. "A simple introduction to Support Vector Machines," 2003.

이 성 욱

e-mail : leesw@cjnu.ac.kr

1996년 서강대학교 전자계산학과(학사)

1998년 서강대학교 컴퓨터학과(석사)

2003년 서강대학교 컴퓨터학과(박사)

2003년~2004년 서강대학교 산업기술연구소 연구원

2003년~2005년 서강대학교 정보통신대학원 대우교수

2004년~2005년 LG전자 기술원 선임연구원

2005년~2007년 동서대학교 컴퓨터공학과 전임강사

2007년~현 재 국립충주대학교 컴퓨터정보공학과 조교수

관심분야 : 형태소 및 구문 분석, 단어의미분별, 대화 언어처리, 한국어 생성 등