Feature Selection을 이용한 SVM 스팸 메일 분석

이광수 손기락
한국외국어대학교 대학원 전자계산교육학과
lgs1010@hanmail.net, ksohn@hufs.ac.kr

Spam mail analysis using SVM with feature selection

Kwang-Su Lee, Kirack Sohn
Computer Education Major,
Hankuk University of Foreign Studies

요 약

오늘날 인터넷 환경의 급속한 발전으로 인하여 이메일을 통한 메시지 교환이 급속히 증가하고 있다. 그러나 이메일의 편리성에도 불구하고 개인이나 회사에서는 스팸 메일로 인한 시간과 비용의 낭비가 크게 증가하고 있다. 이러한 스팸 메일에 대한 문제들을 해결하기 위해서 많은 방법들이 연구되고 있다.

본 논문에서는 패턴 분류 문제에 있어서 우수한 성능을 보이는 SVM(Support Vector Machine)을 사용하여 정상 메일과 스팸 메일을 분류할 수 있는 최적의 항목을 찾고자 한다. 그 방법으로 Feature Selection 기법을 사용하여 항목을 선별하였으며 이 선별된 항목이 얼마나 정확한 구분력을 가지고 있는지를 나타내고자 한다.

1. 서 론

인터넷 포털 업체인 코리아닷컴에 의하면 하루 300~400만통의 메일유통량 중 60% 이상이 광고성 메일로 채워지고 있다고 밝혔다. 이렇게 볼 때, 야후, 라이코스코리아, 마이크로소프트 등 다른 이메일 업체와 일반 기업체의 이메일 서버를 통해 유통되는 분량을 합한다면 하루 유통되는 스팸 메일 건수는 거의 1억통에 육박하고 있는 것이다. 이렇게까지 스팸 메일이 급증하고 있는 데는 이메일 주소가 인터넷 사이트의 게시판이나 방명록 등을 통해 낱낱이 공개되고 있기 때문이다. 결국 스팸 메일이 전체 메일의 절반 이상을 차지하는 수준에 이르다 보니 개인 및 기업에게 스팸 메일 삭제에 엄청난 비용 및 시간의 부담을 준다.

본 논문에서는 패턴 분류 문제에 있어서 우수한 성능을 보이는 SVM(Support Vector Machine)을 사용하여 스팸 메일 여부를 판정하였으며 패턴 분리의 정확성을 증대시키기 위해서 정상 메일(legitimate mail)과 스팸 메일(spam mail)을 분류할 수 있는 최적의 항목(Keyword)을 찾고자 Feature Selection 기법을 사용하여 항목을 선별하였으며 이 선별된 항목이 얼마나 정확한 구분력을 가지고 있는지를 나타내고자 한다.

2. 관련연구

2.1 SVM(Support Vector Machine)

SVM(Support Vector Machine)은 1995년 Vapnik에 의하여 개발되고 제안된 학습 알고리즘이다. 이것은 원래 이진분류(binary classification)를 위하여 개발되었으며 현재에는 생물정보학(bioinformatics), 문자인식, 필기인식, 얼굴 및 물체인식 등 다양한 분야에서 성공적으로 적용되고 있다. 이진분류 문제는 수집된 훈련 데이타를 이용해서 두 클래스를 분류하는 대상함수(target function)를 추정해 내는 과정이라고 볼 수 있다. 그렇게 추정된 분류기는 훈련과정에서 이용되지 않은 새로운 데이타 표본에 대해서도 올바른 결과 값을 낼 수 있는 일반화 성능(generalization performance)이 뛰어나야 한다. SVM은 [그림 1]에서 보는 것과 같이 특징 공간(feature space) 에서 데이타를 나눌 수 있는 초평면(possible hyperplane) 중에서 특정한 초평면(optimal hyperplane)을 선택함으로서 과적합 문제(overfitting)를 방지한다. SVM은 초평면으로부터 가장 가까운 훈련 포이트까지의 최소거리를 최대화시키는 초평면 즉, 최대 여백 초평면(maximum margin hyperplane)을 찾게 된다. SV(Support vector)라고 불리는 두 클래스들 사이에 결정 경계(decision boundary)에 가까이 놓여있는 훈련 예만이 non zero weight를 갖게 된다. SV를 포함하는 초평면 사이의 거리인 여백(margin) 값이 클수록 분류성능은 좋아진다. 이렇게 찾아낸 초평면을 기준으로 테스트를 시행하여 분류 결과를 얻게 된다.

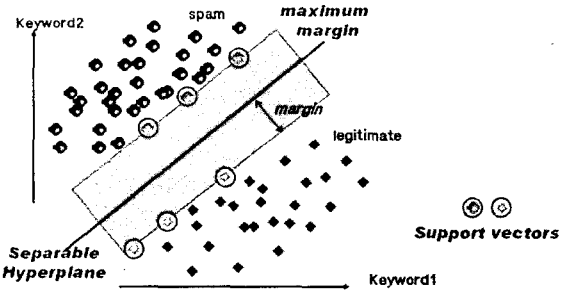

그림 1. Maximum Margin Hyperplane

SVM 이 주목받는 이유는 첫째, 명백한 이론적 근거에 기반하므로 결과 해석이 용이하고, 둘째, 실제 응용에 있어서 인공신경망 수준의 높은 성과를 내고, 셋째, 적은 학습 자료만으로 신속하게 분별학습을 수행할 수 있기 때문이다. 이와 같은 이유로 본 논문에서는 SVM 도구로서 SVMlight를 사용하였다.[1]

2.2 이메일 분석

스팸 메일이 전체 메일의 절반 이상을 차지하는 수준에 이르다 보니 개인 및 기업에게 스팸 메일 삭제에 엄청난 비용 및 시간의 부담을 준다. 따라서 스팸 메일 분류하는 연구가 있어왔다.[2] 본 논문에서는 이런 사항을 개선하기 위해서 정상 메일과 스팸 메일을 분류할 수 있는 최적의 항목을 찾고자 하는데 목적을 두고 있다. 이 메일 분석의 방법으로는 다음의 순서대로 작업을 하였다.

1 단계 : 다수의 사용자 이 메일 계정으로부터 정상 메일과 스팸 메일을 수집한다. [그림 2]

메일 수신자 측면에서 정상과 스팸을 분류하여 정상을 -1로 스팸을 +1로 표시하였다.

```
★[대출] 누구나 안전하고 빠르게::무담보,무보증으로 최고 5,000만원까지 OK!!                +1
★+▶신청하신 분들께 공인중개사 수험자료를 무료로 드립니다!!! 6n1z23                    -1
[설문]설문에 참여하시면 문화상품권을 드립니다!!                                        +1
[학습] Morning Letter (E) - 2005/02/15                                                -1
★고객님은 무방문 대출이 가능합니다.ㅁ                                                  +1
100% 무담보담으로 100% 승인율(최저 이자) 자랑~!★^^★                                    +1
CJ무비패밀리와 함께하는 맛있는 영화<달콤한 인생>시사회에 초대합니다.                   -1
◆농장직송◆허브농장 돌맞이 허브화분 3종세트 특가판매, 고급도자기화분 사용!!            -1
원하는 이상형을 찾아보세요. 193                                                       +1
[중국어열풍]중국어1:1무료학습, 온라인무료강좌                                          -1
직장인 2000만원 누구나 5분대출                                                        +1
```
그림 2. mail data

2 단계 : 수집된 이 메일에서 메일 구분력이 있는 항목 추출한다.

3 단계 : 추출된 항목이 수집된 이 메일에서 표현되는 횟수를 계산 한다.

4 단계 : 수집된 이 메일과 항목을 가지고 Training data와 Test data을 만든다. [그림 3]

```
+1 1:1 23:1
+1 2:1 49:1
-1 3:1
+1 4:1 5:1 51:1
+1 6:1 7:1 13:1
+1 6:1 8:1
+1 6:3 9:1 10:1 11:1 12:2 13:1 52:1
-1 14:1 30:1 65:1 71:1 135:1
-1 15:1
```
그림 3. Training data

5 단계 : Training data에서 SVM Model을 생성한다.

6 단계 : Test data을 사용하여 SVM Model을 평가한다.

Accuracy	85.23%
Precision	94.35%
Recall	86.03%

표 1. SVM 메일 구분 결과

실험을 통해 얻어지는 결과([표 1])를 분석하기 위한 성능 측정의 기준은 아래 [그림 4]에서 보는 것과 같이 accuracy, precision, recall의 세가지 값으로 요약된다.

각 결과값이 가지는 의미는 먼저 accuracy값이 모든 실험에서 SVM이 맞게 예측한 비율을 나타내는 것이고 precision값은 SVM이 스팸이라고 분류한 경우에 실제로 스팸인 비율을 나타내는 것이며, recall값은 실제로 스팸인 경우에서 SVM이 스팸을 예측한 비율을 나타낸 수치이다.

```
- Accuracy  = (tp + tn) / (tp+fp+tn+fn)
- Precision = tp / (tp+fp)
- Recall    = tp / (tp+fn)

tp : True positive
fp : False positive
tn : True negative
fn : False negative
```
그림 4. 각 결과값에 대한 정의

본 논문에서는 precision값이 중요한 역할을 한다. 예를 들어 precision값이 높게 나올시 스팸 메일이 아닌데 스팸 메일이라고 분류하는 오류를 줄일 수 있다. 본 논문의 연구내용도 precision값을 100%에 근접하도록 항목을 선별하는 연구를 하는 것이다.

2.3 Feature Selection

[표 1]은 SVM 메일 구분결과를 나타내고 있다. precision값이 100%에 이르도록 정상 메일과 스팸 메일을 분류할 수 있는 최적의 항목을 찾고자 Feature Selection 기법을 사용하여 항목을 선별하는 것이 이 논문의 주요 내용이다. 마이크로어레이 데이터를 이용한 암세포 분류기법에 사용한 Feature Selection 기법을 응용한다.[3]

분석의 방법으로는 다음의 순서대로 작업을 하였다.

1 단계 : 각 항목의 평균과 표준편차를 이용하여 Feature Selection 기법의 공식을 유도할 수 있다.

$$F = \frac{|\mu^+ - \mu^-|}{|\sigma^+ + \sigma^-|}$$

σ^+ = 정상메일표준편차 $\quad \sigma^-$ = 스팸메일표준편차

μ^+ = 정상메일평균 $\quad \mu^-$ = 스팸메일평균

2 단계 : 유도된 공식 F를 항목별로 내림차순으로 정렬하여

구분력이 없는 항목을 삭제한다.[그림 5]

```
5           0.6449528932571411
58          0.3483664989471436
156         0.3483664989471436
7           0.3023718693527985
50          0.3023718595504761
77          0.2773499488830566
133         0.2635232806205750
157         0.2617850303649902
36          0.2589190900325775
8           0.2472064644098282
95          0.2391824722290039
```

그림 5. 항목별 공식 F의 값

[그림 5]의 항목 값은 정상 메일과 스팸 메일의 구분력을 나타내는 수치이며 이 수치값이 0으로 갈수록 구분력이 없다는 것을 나타낸다. 이 수치값이 0에 가까운 항목을 삭제하고 다시 SVM을 Training하면 [표 2]의 결과값이 나타난다.

Accuracy	86.93%
Precision	98.13%
Recall	85.37%

표 2. SVM 메일 구분 결과

3. 결론 및 향후 연구과제

[표 2]에서 precision값이 [표 1]의 precision값 보다 높게 나온다. 3.78%개선되어 졌다. 이는 Feature Selection기법으로 항목을 선택하면 구분력이 높은 항목을 선택 할 수 있다는 것을 나타내고 있는 것이다.

본 논문에서는 정상 메일과 스팸 메일을 분류할 수 있는 최적의 항목을 찾고자 Feature Selection기법을 제시하였다.

이번 연구에는 항목을 수동으로 추출하는 방법으로 작업하였으나 향후에는 항목을 자동으로 추출하는 방법이 필요하다.

참고문헌

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

[2] 서정우 (2003) n-Gram 색인화와 SVM을 사용한 스팸메일 필터링에 대한 연구. 고려대학교 정보보호대학원 석사학위논문

[3] Terrence S. Furey, *et al.*, Support vector machine classification and validation of cancer tissue sample using microarray expression data, Bioinformatics, Vol.16 no 10 2000, 906-914