Support Vector Machine을 사용한 스팸메일 탐지 방안

서정우° 손태식* 서정택** 문종섭*
*고려대학교 정보보호 기술연구센터, **국가보안기술연구소
{korea002°, 743zh2k, jsmoon}@korea.ac.kr, seojt@etri.re.kr

An Approach for Detecting Spam Mail using Support Vector Machine

Jungwoo Seo° Taeshik Sohn* Jungtaek Seo** Jongsub Moon*
*CIST Korea University, **NSRI

요 약

인터넷 환경의 급속한 발전으로 인하여 전자우편을 통한 메시지 교환은 급속히 증가하고 있다. 하지만 전자우편의 편리성에도 불구하고 개인이나 기업에서는 스팸메일로 인한 시간과 비용의 낭비가 크게 증가하고 있다. 기존의 스팸메일에 대한 연구는 패턴 매칭에 의한 분류나 확률에 의한 분류가 대부분인데, 이와 같은 방법들은 변형된 형태의 메일에 대한 탐지에 있어서 비효율적이다.

본 논문에서는 기존의 연구에 대한 문제점을 보완하기 위하여 패턴 분류문제에 있어서 우수한 성능을 보이는 SVM을 이용하여 정상적인 메일과 스팸메일을 구분하는 방안에 대하여 제시한다.

1. 서 론

인터넷 사용자가 늘어나면서 전자우편 시스템의 사용자는 크게 증가하였다. 전화와 달리 전자우편 시스템을 이용하면 수취인의 부재와 관계없이 메일을 보내거나 송신자가 원할 때에 메일을 보낼 수 있다. 특히, 같은 내용의 메일을 많은 사람에게 동시에 보내는 경우 수신처를 복수로 지정하거나 그룹화 시켜서 발송할 수 있다. 하지만, 웹 페이지의 게시판이나 뉴스그룹에서 획득한 메일 주소 리스트를 이용하여 상업적인 내용이나 원하지 않는 메일을 무차별적으로 발송하는 문제점이 있는데 이를 스팸메일 이라고 한다. 올해의 스팸메일은 1조9천6백억 통에 이를 것으로 전망하고 있으며, 이는 전체메일 중 스팸메일이 차지하는 비중이 40%를 넘을 것으로 추정하고 있다. 이는 지난 2001년 8%에 불과했던 것이 최근 5배나 늘어난 수치이다. 결국, 스팸메일은 개인 및 기업에게 스팸메일 삭제에 엄청난 비용 및 시간의 부담을 준다.

지금까지의 스팸메일 대응 연구는 확률적인 방법[1][2]이나 송신자의 메일주소나 제목, 내용의 특정한 단어들로 정의된 룰 셋의 매칭을 통하여 이루어져 왔다. 하지만, 기존의 스팸메일 대응 방법은 데이터 셋이 증가할 경우 검색 시간의 증가 및 시스템의 자원을 감소시킬 수 있으며, 변형된 형태의 메일이 전송될 경우 이를 효과적으로 탐지하지 못할 수 있다.

그러므로, 본 논문에서는 스팸메일 탐지를 위해서 Support Vector Machine(SVM)을 이용한다. SVM은 1995년 Vapnik에 의해 제안된 Universal Feed Forword 네트워크의 한 종류로서 복잡한 패턴인식과 이진 분류 문제에 있어서 가장 효율적인 해결책으로 알려져 있다[4].

본 논문의 구성은 2장에서 관련된 연구에 대하여 살펴보고, 3장과 4장에서는 SVM의 개요 및 *n*-Gram 기반의 색인 방법에 대하여 알아본다. 5장에서는 스팸메일 탐지 방안에 대하여 알아보고, 6장에서는 본 논문의 실험결과를 설명한다. 마지막으로 7장에서 결론 및 향후 연구 방향을 제시한다.

2. 관련 연구

컴퓨터와 메일서버 사이에서 우편을 관리하기 위한 방법들에 대하여 RFC1725에 규정되어 있다. 스팸메일에 관련한 연구는 인터넷 환경이 발전하면서 더욱 활발하게 수행되고 있다. "A Bayesian Approach to Filtering Junk E-Mail"[2]에서는 Bayesian 이론을 바탕으로 확률적인 방법들에 의하여 분류를 수행하였다. 이렇게 기존의 연구들이 패턴 매칭이나 확률적인 방법[1]들에 많이 의존하고 있다. 하지만 신경망이나 SVM과 같은 비정상 행위 탐지기법은 특정 룰에 의한 탐지인 오용탐지 기법에 비해 변형된 메일이나 알려지지 않은 스팸메일 탐지에 장점을 가진다.

3. Support Vector Machine 개요

3.1 SVM 개요

전통적인 패턴인식 방법들은 경험적인 위험을 최소화하는 반면, SVM은 구조적 위험을 최소화하도록 한다. 패턴 집단이 선형이고 분리 가능한 경우에 있어 SVM은 다음과 같이 설명될 수 있다. 기본적으로 SVM는 입력패턴들을 교사학습 방법을 통하여 +1과 -1의 두 클래스로 패턴을 분류한다. 두 개의 클래스로 분류된 훈련집단은 각 클래스에 포함된 훈련 패턴들을 분리하는 하이퍼플레인(hyperplane)이 결정된다. 하이퍼플레인을 결정하는 입력 패턴들을 support vector라 하며, 적절한 하이퍼플레인을 찾으면 오분류를 피할 수 있다[5].

3.2 분류를 위한 Support Vector Machine

본 절에서는 SVM의 분류와 비선형 함수에 대한 추정에 대한 기본적인 개념을 알아본다[5]. 훈련 데이터 {(xi , di), i = 1, . . . , N}가 주어졌을 때, xi는 두 클래스

중 하나에 속하며, $di \in \{-1, 1\}$는 해당 클래스를 표시하는 라벨의 역할을 한다. SVM은 각 클래스를 구분하는 최적의 분리 경계면을 구하기 위해 분리 경계면과 가장 분리 경계면에 인접한 점과의 거리를 최대화한다. 최적의 선형 분리 경계면을 f(x) = wtx + b로 놓으면, support vector와 f(x)의 거리를 1/||w||로 나타낼 수 있다. SVM은 ||w||2를 최소화하여 분리 간격을 최대화하도록 하여 최적 분리면을 찾아낸다.

커널 함수로는 dot, polynomial, radial, neural과 같은 여러 함수 중 선택할 수 있다.

4. *n*-Gram 기반의 색인 방법

4.1 *n*-Gram 색인방법 개요

본 절에서는 어절 단위 색인법에서 복합명사 띄어쓰기 문제를 완화할 수 있으며, 형태소 단위 해석에서와 같은 복잡한 문장 해석 규칙이나 언어 정보의 개발을 요구하지 않는 색인 방법이다[3].

[표 1] *n*-Gram 기반의 색인 과정

단계 1 : 문서나 질의 내의 모든 어절들을 인식한다.
단계 2 : 불용어를 제거한다.
단계 3 : 각 어절에서 비색인 분절들을 절단한다.
단계 4 : 나머지 색인 분절을 n-Gram들로 분할하여 색인어로 선정한다.

n-Gram 기반의 색인 방법은 검색효과의 측면에서 다음과 같은 장점이 있다. 첫째, *n*-Gram 기반의 색인법은 어절 단위 색인법을 이용할 때의 절단 오류로 인한 파급효과를 완화한다. 둘째, 복합 명사의 띄어쓰기 문제를 완화한다. 셋째, 철자 오류나 일관성 없는 외래어 표기 문제를 적절히 극복할 수 있다.

4.2 *n*-Gram을 이용한 Feature 생성

수집된 메일로부터 획득한 학습 셋이나 테스트 셋은 SVM의 feature로 사용하기 위하여 정규화 과정이 필요하다. 이를 정규화하기 위하여 이전에 정의된 데이터 사전을 활용하여 일정한 크기의 feature를 만들어 낸다. *n*-Gram을 적용하여 feature를 만들어내는 방법은 다음과 같다.

[표 2] *n*-Gram을 적용한 feature 생성

수집된 메일데이터 = {카드증액, 대출 쉽게 하는 방법}
2 Gram 적용데이터={카드, 드증, 증액, 액대, 대출, ... }
3 Gram 적용데이터={카드증, 드증액, 증액대, 액대출, ...}
4 Gram 적용데이터={카드증액, 드증액대, 증액대출, ... }
데이터 사전={카드증액, 연체, 성인광고, 포르노, 대출, ... }
데이터 사전 적용 feature={1, 0, 0, 0, 1, ... }

5. 스팸 메일 탐지 방안

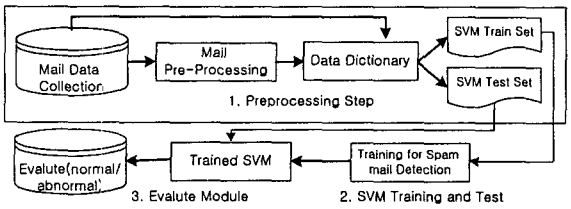
스팸 메일을 탐지하기 위하여 우선 데이터 사전을 생성해야 한다. 데이터 사전 생성방법은 수집된 메일을 분석한 후 대출과 음란성 메일에 사용된 단어의 빈도수를 기준으로 데이터 사전을 생성한다. 이것은 학습데이터나 테스트 데이터의 feature로 사용하기 위하여 정규화 시키는 과정이라고 할 수 있다. [표 3]은 생성된 데이터 사전의 예이다.

[표 3] 데이터 사전

사전 개수	1	2	3	4	5	6	· · · ·	99	100
단어 사전	대출	연체	카드대출	성인광고	누드	대출고민	· · · ·	결제일	대출비법
사전 개수	111	112	113	114	115	116	· · · ·	196	197
단어 사전	노출	쇼걸	포루x	올누드	카드값	카드자금	· · · ·	급한카드	당일해결

생성된 데이터 사전은 일정한 크기의 feature를 생성하게 되는데, 학습 셋과 테스트 셋을 [표 3]의 데이터 사전에 매칭 시킨다. *n*-Gram을 수행한 학습 및 테스트 셋의 메일 데이터와 데이터 사전의 단어가 일치할 경우 해당 위치에 1을 표시하고 일치하지 않으면 0을 표시하면서 feature를 구한다.

SVM을 적용하기 위한 학습 셋이나 테스트 셋이 정의되면 (그림 1)에서와 같이 스팸메일 탐지를 위한 학습을 수행한다. 학습을 수행한 SVM은 테스트 셋에 정의된 메일이 스팸메일인지 여부를 판정하게 된다.



(그림 1) SVM을 사용한 스팸메일 탐지 구성도

6. 스팸메일 탐지 실험 및 결과

6.1 실험 방법

Support Vector Machine을 사용한 스팸메일 탐지를 위하여 학습 데이터 집합과 테스트 데이터 집합 그리고 데이터 사전을 구성해야 한다. 데이터 사전은 학습 및 테스트 데이터의 feature들을 나타내기 위하여 사용되며, 대출관련 메일과 성인사이트 광고를 포함하는 메일을 탐지하기 위하여 구성된다. 데이터 사전의 개수는 197개로 구성되며, 197차원으로 구분된다[표 4 참조].

[표 4] SVM에 적용되는 feature

데이터셋 \ 개수	1	2	3	4	5	· · · ·	196	197
데이터 셋(1)	0	0	1	1	1	· · · ·	0	1
데이터 셋(n-1)	1	1	0	0	1	· · · ·	0	1
데이터 셋(n)	0	0	0	1	1	· · · ·	1	0

스팸메일 탐지를 위한 학습 데이터 집합을 구성해야 하는데, 사용자 메일 서버로부터 수신한 메일을 사용한다. SVM 학습을 위한 데이터는 정상적인 메일 데이터

200개와 스팸메일 200개로 각각 구성된 400개의 메일 데이터를 사용한다[표 5 참조].

[표 5] SVM 학습을 위한 데이터

데이터셋 / 메일종류	학습 데이터 (총 400개)
정상 메일	대출 및 성인 사이트를 제외한 정상 메일(200 개)
스팸 메일	대출 및 성인 사이트를 포함한 스팸 메일(200 개)

실제적으로 스팸메일 여부를 판정하기 위하여 테스트 메일을 구성해야 하는데, 위에서 설명한 학습데이터 수집 방법과 같이 사용자로부터 수집한 메일을 사용한다. 테스트 데이터 구성은 각각 정상메일 500개와 스팸메일 500개로 구성하였다[표 6 참조].

[표 6] SVM 테스트를 위한 데이터

데이터셋 / 메일종류	테스트 데이터 (총 1000개)
정상 메일	대출 및 성인 사이트를 제외한 정상 메일(500 개)
스팸 메일	대출 및 성인 사이트를 포함한 스팸 메일(500 개)

본 실험에서는 [6]에서 개발된 mySVM 공개 라이브러리를 사용하여 실험하였으며, 커널 함수로는 polynomial을 사용하였다. 그리고 Degree는 3으로 고정하여 실험하였다.

6.2 실험 결과

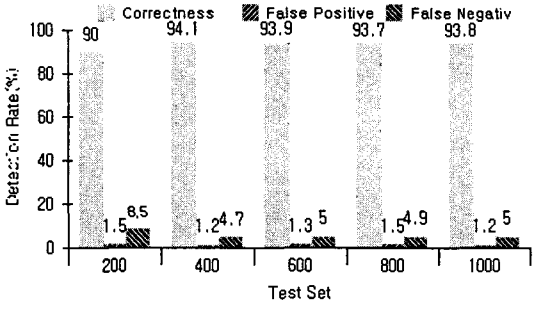
테스트 메일에 대해서 정상메일과 스팸메일을 판정하기 위하여 위에서 정의한 실험방법과 같이 학습 셋에 대하여 SVM에 적용한 커널과 feature의 수 변화에 따른 각 테스트 셋의 탐지 결과를 분석하였다. 학습 셋의 수는 정상메일 200개와 스팸메일 200개를 사용하였으며, 테스트 셋의 수는 정상메일과 스팸메일을 각각 100개씩 증가시키면서 학습을 수행하였다(그림 2 참조). [표 7]에서와 같이 테스트를 위한 학습 셋은 400개이며 테스트 셋은 정상메일과 테스트메일 각각 500개를 사용하였다. 커널함수는 Polynomial을 사용했으며, SVM의 feature는 197차원을 적용시켰다. 테스트의 결과 값은 93.8%의 정상인식률과 1.2%의 False Positive값 그리고 5%의 False Negative값을 확인할 수 있었다.

[표 7] SVM을 이용한 스팸메일 탐지 결과

	커널 함수	Feature (dimension)	테스트 셋 (1000개)		
			FP	FN	TC
학습 셋(400개)	Polynomial	197	1.2	5	93.8

* Ploynomial Kernel의 Degree = 3, FP: False Positive(%), FN: False Negative(%), TC: Total Correctness(%).

(그림 2)에서와 같이 테스트 셋의 메일이 정상메일이지만 스팸메일로 판정하는 False Positive와 스팸메일이지만 정상메일로 판정하는 False Negative의 비율은 메일의 증가와 관계없이 평균적으로 1.4%, 5%를 나타내고 있다.



(그림 2) 학습 셋에 의한 스팸메일 탐지 결과

7. 결 론

본 연구에서는 전자우편을 이용한 스팸메일을 탐지하는 방법에 대한 기존의 연구들을 살펴보고 Support Vector Machine(SVM)을 사용한 새로운 방안을 제시하였다. 제안된 방안은 기존의 확률에 의한 방법이나 룰 셋에 의존하는 방법이 아닌 SVM을 사용함으로서 변형되거나 알려지지 않은 스팸메일 탐지에 있어서 효율적이다. 특히, 데이터 사전으로 Features을 구성함으로서 학습 셋이나 테스트 셋들에 동일한 Feature를 구현하였다. 여기서 구현된 Feature를 이용하여 스팸메일을 탐지하는 방안을 제안하였다.

향후 실험에서는 보다 다양한 형태의 메일을 수집하는 것이 필요하며, 또한 커널의 종류와 테스트 인자들을 다양화하면서 실험을 진행하는 것이 필요하다.

참 고 문 헌

[1] Ion, A., Georgios, P., Vangelis K., Georgios, S., Constantine, D., "Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach", PKDD 2000, pp. 1-13, Sep. 2000.

[2] Mehran, S., Susan, D., David, H., Eric, H., "A Bayesian Approach to Filtering Junk E-Mail", In AAAI-98 Workshop on Learning for Text Categorization, 1998.

[3] 이준호, 안정수, 박현주, 김명호, "한글 문서의 효과적인 검색을 위한 n-Gram 기반의 색인 방법", 정보관리학회지, 1996.

[4] Pontil, M. and Verri, A., "Properties of Support Vect or Machines", A.I. Memo No. 1612; CBCL paper No. 152, Massachusett s Institute of T echnology , Cambridge, 1997.

[5] Cristianini N., Shawe-Taylor J., An Introduction to Support Vector Machines, Cambridge University, 2000

[6] Joachmims T, mySVM - a Support Vector Machine, Univerity Dortmund.