베이지안 네트워크와 멀티 레이어 퍼셉트론을 이용한 모바일 스팸 문자 메시지 필터링 방법

홍승범*, 김문현*
*성균관대학교 정보통신공학부 컴퓨터 공학과
e-mail : falkeadler86@gmail.com, mhkim@ece.skku.ac.kr

A Method for Spam SMS Filtering Using Bayesian Network and Multi Layer Perceptron

Seung-Beom Hong*, Moon-Hyun Kim*
*Department of Computer Engineering
Sungkyunkwan University

요 약

스팸 메시지는 불특정 다수에게 보내지는 광고성 메시지로서 최근 들어 그 양이 증가하고 있는 추세이다. 본 논문에서는 모바일 환경에서의 스팸 메시지 필터링을 위한 시스템을 제안하며 기존 환경에서 자주 사용되었던 키워드 기반 필터링 시스템의 단점을 해결하고자 고안되었다. 베이지안 네트워크를 통해 스팸 메시지들의 패턴을 추출하고 추출된 패턴을 멀티 레이어 퍼셉트론을 이용해 학습하여 메시지들을 분류한다. 이 시스템을 통해 약 93.5%의 필터링 정확도률을 얻었으며 키워드 선택 대신 스팸 메시지를 선택해 학습시킴으로서 사용하기 쉽고 사용자에 맞는 시스템을 구성할 수 있었다.

1. 서론

스팸은 전자 우편, 게시판, 문자 메시지, 전화, 인터넷 포털 사이트의 쪽지 기능 등을 통해 불특정 다수의 사람들에게 보내는 광고성 편지 또는 메시지를 말한다. 이러한 스팸 메시지는 그 양이 점점 줄어왔으나 최근 들어 다시 증가하고 있는 추세이다.[1] 초기에는 스팸 메일에 대해 나이브 베이지안 규칙을 이용한 필터링[2], 메일의 메타정보를 신경망으로 학습하는 필터링 방법[3] 등이 자주 사용되었다. 현재는 휴대폰 사용량의 증가로 문자메시지 또한 통신수단으로서 폭넓게 사용되고 있다. 이와 함께 메일 시스템에서 문제가 되었던 스팸 메일이 그 범위를 넓혀 모바일의 영역까지 확대하고 있다. 앞의 필터링 방법들은 데이터베이스의 크기, 필요한 메타정보 특성, 컴퓨팅 파워 등의 특성으로 인해 모바일 환경에서 사용되기에는 제약이 있다. 현재 모바일 환경에서 사용되는 필터는 일반적으로 키워드 기반 방법 또는 정규 표현식을 이용한 방법으로서 False Positive의 비율이 높고 사용자가 필터링할 키워드를 입력해야 하는 불편함이 있다. 이에 본 논문에서는 베이지안 네트워크를 이용해 스팸 메시지들의 공통된 패턴을 추출하고, 이 패턴을 멀티 레이어 퍼셉트론을 이용해 학습시켜 새로운 문자 메시지가 수신되었을 때 학습된 멀티 레이어 퍼셉트론을 이용해 스팸 여부를 판단하는 시스템을 제안함으로서 키워드 기반 시스템의 단점인 False Positive의 비율을 줄이고, 키워드를 입력해야 하는 과정을 스팸 메시지의 선택 및 학습으로 대체함으로서 사용하기 편리하고 사용자에 맞는 시스템을 구현하고자 한다.

2. 관련 연구

2.1 K2 알고리즘

스팸 문자메시지의 패턴을 추출하기 위해서는 베이지안 네트워크 그래프가 필요하다. 따라서 본 논문에서는 베이지안 네트워크의 학습방법으로 K2 알고리즘을 사용하였다. K2 알고리즘은 Gregory F. cooper와 Edward Herskovits에 의해 제안된 베이지안 네트워크 학습 알고리즘으로서 주어진 데이터에 대한 가장 높은 사후 확률을 가질 수 있는 네트워크를 찾는 방법이다. K2 알고리즘은 그래프를 구성하는 노드들의 순서를 미리 결정하여 각 노드가 가질 수 있는 부모 노드를 노드 순서에서 선행한 노드들로 제한하고, 최대 부모의 개수를 지정하여 학습공간을 줄인다. 각 노드가 가질 수 있는 모든 부모들과의 관계를 나타내는 K2 메트릭 점수를 이용해 가장 큰 값을 가지는 경우를 선택하여 네트워크를 구성하는 탐욕적인 방법을 사용한다. 각 노드의 K2 메트릭 점수를 계산하는 평가함수 G 함수의 핵심 수식은 다음과 같이 계산한다.

$$P(G,D) = P(G)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\prod_{k=1}^{r_i}N_{ijk}! \qquad (1)$$

다음은 표 1은 K2 알고리즘의 탐색 과정이다. [4]

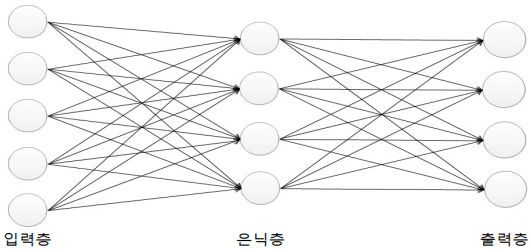
<표 1> K2 알고리즘 Pseudo Code

```
for I := 1 to n do
    π_i := 0;
    P_old := g(i,π_i);
    OKToProceed := true;
    while OKToProceed and |π_i| < u do
        let z be the node in Pred(x_i) - π_i
        P_nw := g(i,π_i ∪ {z});
        if P_nw>P_old then
          P_old := P_nw;
          π_i := π_i ∪ {z};
        else OKToProceed := false;
    end {whlie};
    write('Node:',x_i, 'Parents of this node:',π_i)
end{for};
```

Input : node set(n), order, upper bound(u), DB(d)
Output : Node, parents nodes

2.3 패턴 분류기로서의 멀티 레이어 퍼셉트론

신경망은 생명체의 신경조직에서 착안하여 모델화한 정보처리 시스템으로서 외부로부터 받아들이는 입력에 대해 동적반응을 일으킴으로서 필요한 출력을 생성하며 주어진 예제 패턴을 반복 학습하여 스스로 지식을 획득하고, 오류 내성, 일반화, 적응성 등의 특징 때문에 패턴 인식에 있어서 적합하다. 멀티 레이어 퍼셉트론은 입력 층, 출력 층, 하나 이상의 은닉 층을 포함하는 구조를 가진다.



(그림 1) 일반적인 멀티 레이어 퍼셉트론의 구조

이러한 멀티 레이어 퍼셉트론은 각 유닛의 입출력 특성함수를 시그모이드 비선형 활성함수로 이용하면 미분이 가능해 은닉층을 학습시킬 수 있는 역전파 알고리즘을 사용할 수 있다. 역전파 알고리즘은 입력 층부터 출력 층까지 각 층의 결과 값으로 다음 층의 출력을 계산하고 목표 출력 값과의 오차를 계산하여 역방향으로 연결강도의 오차를 수정해 나가며 학습한다.

2.4 나이브 베이지안 스팸 필터링

폴 그래이엄은 기존의 규칙기반 필터링에서 벗어나 스팸을 학습하고 테스트하는 과정을 거쳐 메시지가 스팸일 확률이 얼마나 되는 가를 구해 분류하는 방법을 제안했다.[2] 스팸 문서와 스팸이 아닌 문서집합을 이용하여, 각 집합의 문서들을 토큰화하여 토큰들의 등장 횟수를 기반으로 각 토큰이 스팸 메시지에서 스팸 토큰으로서 사용된 확률을 계산한다. 새로운 메시지가 도착하면 토큰화한 후 토큰들이 가지는 확률을 비교하여 0.5에서 차이가 큰 순서대로 15개를 선택하고 각 토큰들의 확률을 이용해 메시지의 스팸 확률을 식 2로 결정한다.

$$P(S|W_1, W_2, \ldots\ldots W_{15}) \qquad (2)$$

15개의 토큰에 대해서 각 토큰이 나온 경우에 해당 메시지가 스팸인 확률을 구하는 것이다. 여기서 스팸이 아닐 경우 토큰 등장 확률은 식 3으로 계산한다.

$$P(W_n|H) = 1 - P(W_n|S) \qquad (3)$$

나이브 베이지안 가정을 이용해 식 4와 같이 계산한다.

$$\frac{P(W_1|S)\ldots\ldots\ldots P(W_{15}|S)}{P(W_1|S)\ldots\ldots\ldots P(W_{15}|S) + P(W_1|H)\ldots\ldots\ldots P(W_{15}|H)} \qquad (4)$$

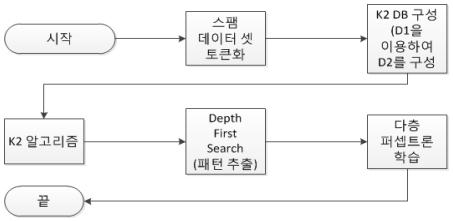
3. 베이지안 네트워크와 멀티 레이어 퍼셉트론을 이용한 스팸 필터링 시스템

3.1 스팸 필터링 시스템의 구성

스팸 필터링 시스템은 두 단계로 진행된다. 학습단계에서는 K2 알고리즘을 이용해 베이지안 네트워크를 구성하여 패턴을 추출하고, 멀티 레이어 퍼셉트론을 이용해 패턴을 학습한다. 베이지안 네트워크 특성인 변수들 간의 관계를 표현하는 점을 이용하여 토큰 사이의 선후 관계를 찾아내고, 관계에 포함된 토큰들을 하나의 패턴으로서 사용했다. 이 패턴을 멀티 레이어 퍼셉트론으로 학습시키며 멀티 레이어 퍼셉트론의 일반화 특성을 이용하여 학습된 멀티 레이어 퍼셉트론을 분류기로서 사용한다.

3.1 필터링을 위한 학습단계

본 절에서는 스팸 메시지를 구성하는 각 토큰들의 위치와 토큰간의 관계를 이용해 분류기준 학습방법을 설명한다. 학습 시스템은 2개의 데이터베이스DB1, DB2와 이용해 스팸 패턴 그래프를 구성한 후 한 개의 패턴 데이터 셋을 추출한다. 추출된 패턴 셋을 이용해 멀티 레이어 퍼셉트론을 학습시킨다. 그림 2는 학습 단계의 순서도이다.

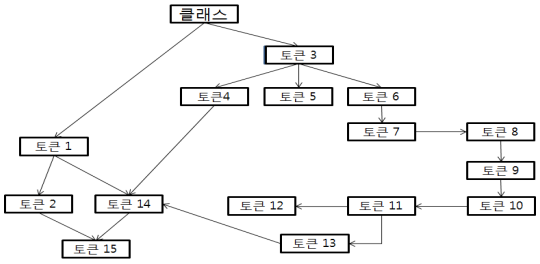


(그림 2) 학습단계의 순서도

학습 단계에서는 스팸 메시지를 구성하는 토큰들을 이용한다. 먼저 각 토큰들의 스팸 확률을 가지고 있는 데이터베이스 DB1이 필요하다. DB1은 다수의 스팸 메시지와 비스팸 메시지를 토큰화하고, 각 토큰이 스팸 메시지에서 사용된 확률을 저장한 데이터베이스이다. 또한 토큰화 된 스팸 메시지에서 패턴을 추출하기 위해 이용할 베이지안 네트워크를 생성하기 위한 데이터베이스 DB2가 이용된다. DB2 데이터베이스는 하나의 레코드가 15개의 토큰 변수와 하나의 클래스 변수를 가진다. 규칙은 식 5와 같다.

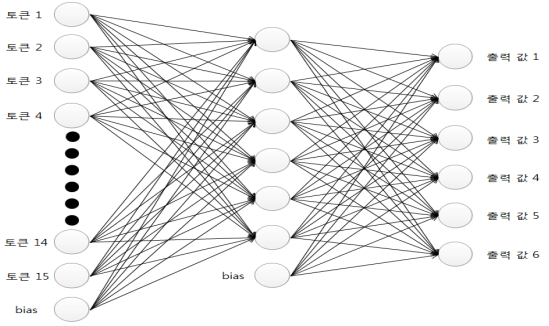
$$Pattern(T) = \begin{cases} 1, & \text{if } P(T) \geq 0.5 \\ 0, & otherwise \end{cases} \quad (5)$$

식 5에서 T는 각 토큰을 의미하며 P(T)는 해당 토큰의 스팸 확률을 의미한다. 토큰들은 원래 메시지에서 등장하는 순서대로 정렬하며 15개의 토큰을 이용한다. 따라서 15개 이상이 되는 토큰은 제거하고, 15개 미만의 토큰을 가지는 메시지는 NULLSTRING으로 표시되는 더미 토큰을 추가하여 15개를 맞춘다. 여기서 DB1이 사용되는데 DB2의 각 토큰 변수의 토큰들의 확률 값이 0.5이상이 되면 1로, 0.5미만이거나 더미토큰인 경우 0으로 매칭 하여 데이터베이스에 저장한다. 이렇게 구성한 DB2를 가지고 K2 알고리즘을 이용해 베이지안 네트워크를 구성한다. 학습 시 노드 오더는 토큰이 등장하는 순서대로 결정하며 학습된 베이지안 네트워크는 스팸 패턴 그래프라고 말한다. 그림 3은 스팸 패턴 그래프의 예제이다.



(그림 3) 구성된 스팸 패턴 그래프의 예

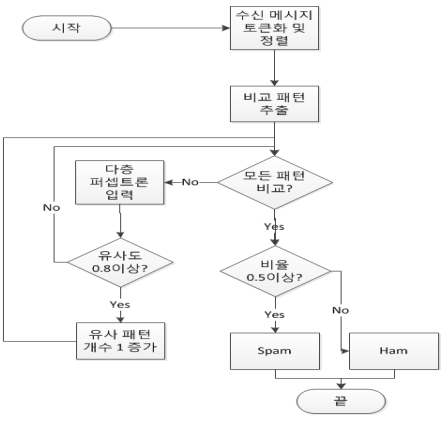
학습된 네트워크에서 Depth First Search를 이용하여 패턴을 추출한다. 베이지안 네트워크는 Directed Acyclic Graph 의 특성을 가지고 있기 때문에 하나의 트리로 생각할 수 있으며 트리내의 각 경로는 토큰들 간의 부모 자식 관계를 나타낸다. 따라서 트리의 루트부터 리프까지의 하나의 경로는 스팸 문자들이 가지는 토큰들의 관계를 나타내는 패턴이라고 생각할 수 있다. 또한 그려지는 스팸 패턴 그래프의 그래프는 노드 개수가 16개로 고정되어 있다. 따라서 Depth First Search를 통해 항상 리프 노드에 도달 할 수 있다. 추출된 패턴은 멀티 레이어 퍼셉트론을 이용해 학습한다. 멀티 레이어 퍼셉트론은 하나의 입력 층, 하나의 은닉 층, 그리고 하나의 출력 층의 3개 층으로 구성된다. 입력 층은 15개의 토큰 노드, 1의 개수를 입력하는 노드, bias 노드의 17개로 구성되며, 출력 층은 패턴의 개수에 따라 다르게 생성한다. 입력 층의 1 에서 15 노드는 수신된 문자 메시지를 토큰화해 얻어진 토큰 15개를 이용한다. 또한 패턴의 1은 스팸 토큰을 의미하는데 멀티 레이어 퍼셉트론은 1의 개수를 나타내는 특성 값이 없을 경우 0으로 학습할 수 있기 때문에 1의 개수를 특성 값으로 활용해 입력층의 마지막 노드에 추가하여 패턴을 학습하도록 하였다. 각 뉴런의 활성함수는 시그모이드 함수를 사용하고, 역전파 알고리즘을 사용하여 전체 네트워크 에러가 0.01이하가 될 때 까지 반복하여 패턴을 학습한다. 그림 4는 멀티 레이어 퍼셉트론의 구조이다.



(그림 4) 구성된 멀티 레이어 퍼셉트론

3.2 스팸 토큰간의 관계를 이용한 분류 방법

본 절에서는 학습된 멀티 레이어 퍼셉트론을 이용해 스팸 메시지를 분류하는 방법을 설명한다. 학습된 멀티 레이어 퍼셉트론은 새로운 메시지가 수신될 때 분류기로서 동작한다. 메시지 수신 시 처리 과정은 그림 5와 같다.



(그림 5) 스팸 여부를 판단하는 순서도

수신된 메시지를 데이터베이스 DB2를 구성 할 때와 마찬가지 과정으로 변환한다. 다음은 각 패턴들과 정확히 비교하기 위해 입력 데이터에서 패턴 정보를 추출할 필요가 있다. 매칭 까지 끝마친 메시지는 여러 개의 패턴들을 동시에 가지고 있을 수 있기 때문에 정확한 패턴 간의 비교를 위해서 현재 입력 값이 가지고 있는 각 패턴의 형태를

추출해서 따로 비교해야 한다. 여러 개의 패턴을 동시에 가지고 있는 경우, 경험적으로 멀티 레이어 퍼셉트론이 제대로 분별 해 낼 수 없는 경우가 있었다. 각 패턴의 형태를 입력 값에서 추출해 내는 방법은 표 2와 같다. 표 2의 맨 윗줄은 입력 값이고, 그 아래 줄은 하나의 특정 패턴이다. 이 패턴과 입력 값을 토큰별로 &연산하여 추출한다.

<표 2> 입력 값에서 패턴 형태를 추출하는 방법

0	0	1	0	0	1	1	1	1	1	1	0	1	1	1
&	&	&	&	&	&	&	&	&	&	&	&	&	&	&
1	0	1	1	0	1	0	0	0	0	1	0	1	0	1
0	0	1	0	0	1	0	0	0	0	1	0	1	0	1

여기서 중요한 것은 1이므로 패턴에 존재하는 1의 위치에 실제 입력 값이 1을 가지고 있는지를 나타낸다고 할 수 있다. 이렇게 패턴을 이용해 추출된 형태의 입력 값과 1의 개수를 멀티 레이어 퍼셉트론에 입력하면 학습된 정보를 바탕으로 출력을 계산하며 출력 값은 각 패턴과의 유사도를 나타낸다. 각 패턴 입력에 대한 출력 노드의 결과가 0.8이상인 경우 해당 패턴과 비슷하다고 판단한다.

$$Result(Sp, Tp) = \begin{cases} Spam, & \text{if } Sp/Tp \geq t \\ Ham, & otherwise \end{cases} \quad (7)$$

식 7에서 Sp는 결과가 0.8이상인 패턴의 수이고, Tp는 전체 패턴의 수, t는 사용자가 지정한 Threshold다. 식의 결과가 Threshold 이상이면 스팸으로 결정한다.

3.3 실험 결과

학습 데이터는 46개의 스팸, 59개의 비 스팸 메시지를 이용해 토큰 확률을, 46개의 스팸 메시지를 이용해 스팸 패턴 그래프를 학습했다. 추출된 패턴은 표 3과 같다.

<표 3> 스팸 패턴 그래프를 통해 추출된 패턴

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	0	0	0	0	1	1
0	0	1	1	0	0	0	0	0	0	0	0	0	1	1
0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	1	1	1	1	1	1	1	0	0	0
0	0	1	0	0	1	1	1	1	1	1	0	1	1	1

패턴을 이용해 전체 네트워크 에러가 0.01이하가 될 때까지 신경망을 학습시켰으며 학습 횟수는 평균 520회이다. 테스트 데이터는 87개의 스팸 메시지, 65개의 비 스팸 메시지를 이용했다. Threshold는 0.5로 설정했다. 테스트 결과는 혼동행렬을 이용해 나타내며, 그 결과는 표 4와 같이 나타난다. 이 결과를 이용해 성능을 평가한다. 전체 오차율을 식 8을 이용해 계산한다.

<표 4> 스팸 필터링의 결과

	Yes	No
스팸	77(TP)	10(FN)
비 스팸	0(FP)	65(TN)

$$Total\ Error = \frac{(FP+FN)}{(TP+TN+FP+FN)} \times 100 \quad (8)$$

계산 결과 전체 오차율은 6.57%로서 총 93.43% 정확성을 보여주었다. 이 가운데 False Negative는 전혀 학습되지 않은 새로운 형태의 스팸 메시지에서 주로 발생하였다. 특히 False Positive는 0%의 확률을 보였는데 이는 스팸 패턴만을 학습시킴으로서 패턴에 맞지 않는 형태의 문자 메시지는 정상으로 처리했기 때문이라고 생각된다.

4. 결론

본 논문은 키워드 기반 스팸 문자 필터링의 한계를 극복하기 위해 진행되었다. 스팸 메시지들의 패턴을 추출하기 위해 스팸 패턴 그래프를 이용해 패턴을 추출하여 멀티 레이어 퍼셉트론으로 수신된 메시지의 스팸 여부를 판단한다. 스팸 메시지 학습을 통한 필터링 시스템을 이용함으로서 False Positive의 비율을 줄일 수 있었으며, 스팸 메시지만을 선택함으로서 간단하게 필터를 사용자에 맞게 구성하고 사용할 수 있었다. 현재 시스템은 모바일 문자 메시지에 대해서만 적용되지만, 스팸 메일의 특성을 이용해 처리할 수 있는 시스템으로 확장할 수 있다. 앞으로는 스팸 메일을 분류할 수 있는 시스템 을 연구하고자 한다.

참고문헌

[1] 통계청. 1인 1일 스팸 메시지 수신량, http://www.index.go.kr/

[2] Paul Graham, “A Plan for Spam“, Hackers & Painters Big Ideas from the Computer Age, O'Reilly Media, pp.121 - 129, 2004.

[3] I. Stuart, S. Cha, and C. Tappert, “A Neural Network Classifier for Junk E-Mail,” in Document Analysis Systems, 2004, pp. 442-450.

[4] Gregory F. cooper, Edward Herskovits, “A Bayesian Method for the Induction of Probabilistic Networks from Data”, Machine Learning, 9 ,pp. 309-347, 1992.

[5] Andrej Bratkom Gordon V. Cormack Bogdan Filipic. Tomas R. Lynam, Blaz Zupan, “Spam Filtering Using Statistical Data Compression Models”, Journal of Machine Learning Research, 7, 2006, 2673-2698, 2006

[6] Sadegh Kharazmim, Ali FarahmandNejad, “Spam Email Detection Using Bayesian Spanning Tree”, COMPUTERS and SIMULATION in MODERN SCIENCE, Volume 1, 172-175, 2008