經營論叢 第20輯 第2號(通卷 24號, 2002年 12月)
Ewha Management Review Vol. 20 No.2(Dec. 2002)

# 데이터 마이닝 기법을 활용한 스팸 메일 분류 및 예측모형 구축에 관한 연구

# Spam Mail Classification Modeling Using Data Mining Techniques

신 경 식* · 안 수 산**
Kyung-Shik Shin · Su-San Ahn

〈목 차〉



<초록>

스팸 메일이란 수신자의 의사와 관계없이 불특정 다수에게 무작위로 발송되는 광고성 이-메일을 일컫는 말로, 벌크(bulk)메일, 정크(junk)메일, 언솔리시티드(Unsolicited)메일과도 유사한 의미로 사용된다. 스팸 메일은 비용이 저렴하고 단순하기 때문에 사업자들은 마케팅 수단으로 선호하지만, 사용자들은 이러한 스팸 메일로 인해 정신적, 물리적 스트레스를 받을 뿐 아니라, 이를 송수신하는 과정에서 사용되는 메일 서버에 과도한 부하를 준다. 또한 공공의 성격을 지니는 네트워크 자원을 아무런 비용의 지불 없이 독점하는 등 사회 전반에 걸쳐 상당한 피해를 주고 있다.

본 연구에서는 데이터 마이닝의 기법 중 분류 문제(classification task)에 많이 사용되는 인

* 이화여자대학교 경영대학 교수
** 이화여자대학교 경영대학

공신경망(artificial neural networks)과 의사결정나무(decision tree)기법을 이용하여 스팸 메일의 분류와 예측을 위한 모형을 구축한다.

주제어 : 데이터 마이닝, 인공신경망, 의사결정나무, 스팸 메일, 분류, 예측

## Ⅰ. 서론

정보화 사회가 가져다 준 가장 큰 선물 중 하나로 이-메일(e-mail)을 꼽는다. 일반인들은 이메일을 편지나 전화를 대체하는 정도의 도구로 활용하지만 기업 환경에서 이-메일 사용은 업무공간의 극복, 사내 커뮤니케이션의 극대화 등의 장점을 제공하면서 회사 내 업무흐름을 전체적으로 변화시키는 혁명을 이끌어왔다.

그러나 최근 중요 사회문제가 되고 있는 스팸 메일(spam mail)의 등장은 이-메일의 편리성에 커다란 반대급부를 제공하고 있다. 스팸 메일이란 수신자의 의사와 관계없이 불특정 다수에게 무작위로 발송되는 광고성 이-메일을 일컫는 말로, 벌크(Bulk)메일, 정크(Junk)메일, 언솔리시티드(Unsolicited)메일과도 유사한 의미로 사용된다(Cranor and LaMacchia, 1998).

스팸 메일은 사용자들 개인에게 정신적, 물리적 스트레스를 줄 뿐 아니라, 조직의 업무진행을 지연시키며, 나아가 공공자원인 네트워크 독점에 의한 피해 등 사회 전 분야에 걸쳐 부정적인 영향을 미치고 있다.

본 연구에서는 스팸 메일의 정의와 유형, 문제점, 사회적 문제 등에 관해 살펴보고, 데이터 마이닝 기법을 활용하여 스팸 메일을 사전에 분류할 수 있는 필터링(filtering) 모형의 구축방법론을 제시한다.

본 연구의 구성은 다음과 같다. 2장에서는 스팸 메일의 정의 및 유형, 그리고 문제점들을 중심으로 살펴보고, 3장에서는 인공신경망과 의사결정나무 분석을 중심으로 모형구축을 위한 데이터 마이닝 방법론에 관해 설명한다. 4장에서는 본 연구에서 제시하고 있는 스팸 메일 분류예측 모형에 관해 설명하며, 5장에서는 모형의 성과를 제시하고, 분석해 본다. 마지막으로는 결론과 본 연구의 의의 및 한계에 관해 논한다.

# Ⅱ. 스팸 메일의 정의와 유형, 문제점

## 1. 스팸 메일의 정의

스팸 메일이란 수신자의 의사와 관계없이 불특정 다수에게 무작위로 발송되는 광고성 이-메일을 일컫는다. 벌크(Bulk)메일, 정크(Junk)메일, 언솔리시티드(Unsolicited)메일과도 유사한 의미로 사용된다(Cranor and LaMacchia, 1998).

스팸이 본격적으로 시작된 것은 1995년 4월, 미국에서 법률사무소를 운영하던 로렌스 켄트와 마사 시걸부부가 '매스포트'라는 프로그램을 작성하여 5~6000개에 달하는 뉴스그룹 게시판을 통해 그린카드를 얻고자 하는 외국인을 상대로 하는 광고성 문구를 무작위로 포스팅하기 시작한데서 비롯되었다. 일종의 동호회 성격을 가진 뉴스그룹의 참여자들은 이들의 무단침략행위에 대해 극도의 분노를 표하기 시작했고, 이에 항의하는 메일이 쏟아졌다. 이렇게 쏟아지는 메일로 ISP(Internet Service Provider)회사의 컴퓨터는 다운될 지경에 이르렀고, 그들에게 계정을 제공했던 인터넷 다이렉트사는 그들의 계정을 과감히 삭제해 버렸다(Cranor and LaMacchia, 1998).

이후 미국의 스팸의 황제라 불리우는 사이버프로모션사의 회장인 스탠포드 왈라스는 PC통신 이용자들의 전자우편주소를 데이터베이스로 만들어 광고주에게 판매하였고, 1996년 4월 AOL은 스팸 메일의 60%가 왈라스가 판매한 목록에서 비롯됐다며 왈라스를 미 법원에 고소하였다. 1996년 11월 미국 뉴욕의 연방법원은 AOL의 손을 들어주었고 왈라스는 패소했다(정재윤, 2000).

스패머(Spammer: 스팸 메일을 보내는 사람)들은 주로 ISP나 PC통신 아이디를 무작위로 스패밍한다. 스팸 메일은 많은 사용자들에게 스트레스의 요인이 되고 있다. 스팸 메일 하나를 읽는데 비록 몇 초 밖에 소요되지 않는다고 해도 타인의 시간을 반강제적으로 빼앗는 것이나 다름없고 누적되면 상당한 시간적, 비용적 손실을 가져오게 된다. 최근에는 스팸 메일을 발송할 목적을 가진 사람들을 위한 이-메일 주소 수집기와 같은 소프트웨어까지 등장하고 있으며, 메일 바이러스를 타고 유포되는 경우까지를 포함할 경우 스팸 메일의 문제점은 무시할 수 없는 정도에 이르고 있다(Gillin, 1999). 또한 스팸 메일을 발신하고 수신하는 과정에서 이용되는 메일 서버에 엄청난 부하를 주고 공공의 성격을 지니는 네트워크 자원을 아무런 비용의 지불 없이 독점하게 되는 좋지 않은 결과를 가져온다(정재윤, 2000).

## 2. 스팸 메일의 유형

### 가. 인커밍 스팸(Incoming Spam)

네티즌들이 일반적으로 받아보는 메일 유형으로, 개인 메일계정 혹은 해당 개인이 속한 메일 서버로 보내지는 스팸 메일들을 말하며 트래픽으로 인한 회선비용의 증가와 폭탄메일로 인한 서버 다운현상들이 나타날 수 있다. 또한, 불건전 정보들의 남발에 의한 특정 개인과 해당 개인이 운영하는 사이트의 이용자들에게 피해가 돌아간다.

### 나. 릴레이 스팸(Relay Spam)

인커밍 스팸보다 심각한 경우로서 스패머(spammer)가 자신의 메일 서버가 아닌 상대방이 속한 메일 서버를 이용하여 제 3자에게 스팸 메일을 발송하는 경우이다. 이 경우는 인커밍스팸의 피해뿐만 아니라 상대방이 속한 그룹이 스패머라고 오해를 받고 인터넷 서비스 업체(ISP)들로부터 메일 발송을 거부당하기도 한다.

## 3. 법적 규제와 문제점

스팸에 대한 찬반논쟁은 국내 뿐 아니라 전 세계적으로 활발히 전개되고 있다. 개인의 프라이버시를 중요하게 여기는 이들은 스팸에 대한 강력한 규제를 요구하고 있고, 상업성 광고를 위해 주로 스팸 메일을 사용하는 온라인 사업자들은 스팸에 대한 규제가 인터넷 상거래를 위축시킬 것이라고 맞서고 있다(한국경제, 1999).

미국에서는 수정헌법 1조와 관련하여 논란이 분분한 반면, 국내에서는 1999년 7월, 스팸 메일 발송에 대해 500만원의 벌금을 부과하는 법안이 통과되었다(정보통신 이용에 관한 법률 제 32조).

> 제 32조[과태료] 불법 또는 부당한 방법으로 보호조치를 침해하거나 훼손하는 경우에 해당자는 500만원의 과태료에 처한다.

스팸 메일의 폐해가 심각해져감에 따라 2000년 7월 미국 연방정부는 18개 주에서 2년간에 걸쳐 만든 초안에 따라 광고임을 밝히지 않은 광고메일, 인터넷서비스업체의 정책에 위배되는 메일, 정보제공을 위장한 허위메일, 승낙 없이 스팸 메일을 발송하기 위해 다른 도메인을 사용한 경우, 적발되면 메시지 당 50달러, 하루 500달러, 최고 2만 5000달러의 벌금을 물도록 하는 법안을 통과시켰다. 그러나 워싱턴주와 캘리포니아주에서는 반스팸 메일법이 상거래 활동관련법에 저촉된다는 위헌결정이 나오기도 해서 주마다 시행의 차질을

빗고 있다(http://www.panix.com).

미국의 의회에서 일고 있는 스팸규제 법안은 스팸을 전면 금지하는 Netizen Protection Act 법안(뉴저지주), 스팸은 허용하되 출처와 광고라는 사실을 명시하라는 법안(알래스카주), 스팸업체들에 대한 엄격한 관리를 요구하는 전자메일상자보호법안(뉴저지주) 등 다양하다(http://www.cauce.org).

국내외적으로 법적인 규제를 시행하고 있지만, 법조항의 허점을 이용하는 경우들이 많아 현실적인 적용은 제대로 되지 않고 있는 실정이다. 국내의 경우 비영리 목적의 광고성 메일에 대해서는 전산망법으로는 어떠한 제재도 가할 수가 없도록 되어 있다. 예를 들어 개인이 비영리적인 목적으로 개인홈페이지를 홍보하기 위해서, 혹은 정치적인 목적으로 선거운동을 하기 위해서 무작위로 뽑은 이메일 주소로 메일을 보낸다고 할 때 이는 일반 네티즌에게는 스팸 메일로 인식되며, 개인 정보유출, 사이버상의 개인적 영역의 침해 등 스팸이 가지는 여러 가지 해악을 갖고 있지만 현재 전산망법으로는 스팸으로 규정하지 않는다.

또한 수신자가 거부의사를 보내기 전에는 스팸을 동일인에게 계속 보내도 아무런 문제가 되지 않는다고 명시되어 있다. 전산망법은 스팸을 명문으로 금지하고 있지만 스팸의 정의를 잘못 규정하고 있고 수신자가 거부의사를 보낸다고 하더라도 스팸 발송자는 처음 1회는 아무런 제약 없이 보낼 수 있기 때문에 스팸 발송자로서는 스팸발송의 효과를 충분히 볼 수 있다(정재윤, 2000).

## 4. 스팸막기

스팸 메일을 통한 피해가 국내외 곳곳에서 발견되자, 이를 기술적으로 막는 많은 수단들이 등장했다. 전자우편 서버인 SMTP의 환경설정에는 이런 스팸을 막을 수 있는 기능이 기본으로 내장되기 시작했다(한국경제, 1999).

실제로 국내외 몇몇 ISP들은 자체 기술로 SMTP서버에 스팸방지 기능을 장착하였으며, 네티즌들이 많이 이용하고 있는 유도라와 같은 메일 클라이언트도 필터 기능이 내장돼 이런 광고 편지들을 자동으로 걸러낼 수 있도록 하고 있다. SMTP서버와 메일 클라이언트 프로그램에서의 필터링 기능은 점차 확대되고 있는 추세이다. 특히 대표적인 웹브라우저인 넷스케이프와 MS 익스플로러에서는 기본적으로 필터링 옵션이 장착되어 있다(경영과 컴퓨터, 1999).

1996년부터 스팸을 극복하기 위한 새로운 방법론이 등장하기 시작했는데, 대표적인 것이 Opt-In(선참여) 방식과 Opt-out(선택적불참)방식이다.

먼저 Opt-Out방식이란 광고업자들이 그들의 리스트에서 이름을 지울 기회를 제공하면서 보다 더 주의를 기울여 소비자의 신뢰를 확보하자는 것이다. 그러나 받는 사람의 입장

에서는 왜 무작위로 보낸 메일에 시간과 노력을 들여 해지응답을 해야만 취소되는가에 대해 의문을 갖게 된다. 반면, Opt-in방식은 원하는 정보를 선별하여 이-메일로 받는 방식이다. Opt-in방식의 이-메일 비즈니스는 두 가지 정도로 나눌 수 있다. 첫째는 소비자나 혹은 ISP에게 일정한 비용을 지불하게 함으로써 자신에게 맞는 양질의 정보를 사는 유료화 패턴이다. 이러한 예로는 미국에서는 '데일리프로', '뉴스바이트' 등이 있다. 두 번째 방법은 무료로 정보를 제공해주고 메일이나 게시판에 광고 공간을 허락 받아서 리서치 응답 등을 제공받는 광고형의 패턴이다. 여기서 중요한 것은 단순히 1회성의 광고를 보내서 얻는 효과보다 장기적으로 양질의 콘텐츠를 제공함으로써 얻게 되는 신뢰성의 제고이다(정재윤, 2000).

스팸으로부터 네티즌을 보호하려는 노력이 활발히 전개되고 있지만 시간과 공간을 초월한 대량 살포라는 기술적 장점과 익명성이라는 문화적 특성 때문에 완벽하게 막는다는 것은 현실적으로 불가능하여, 여전히 많은 인터넷 사용자들은 스팸 메일로부터 시달리고 있다.

## Ⅲ. 데이터 마이닝 기법

데이터 마이닝이란 대용량의 데이터로부터 이들 데이터 내에 존재하는 관계, 패턴, 규칙 등을 탐색하고 찾아내어 모형화함으로써 유용한 지식을 추출하는 일련의 과정들을 말한다. 주로 분류(classification), 측정(estimation), 시계열 예측(time-series prediction), 군집화(clustering) 등의 문제에 많이 활용되고 있다. 스팸 메일 필터링 모형의 경우 결국 "분류" 문제이기 때문에 데이터 마이닝 기법의 활용이 가능하다. 본 장에서는 분류문제의 해결에 용이한 데이터 마이닝 기법들인 인공신경망(artificial Neural networks)과 의사결정나무(decision tree) 에 대하여 설명한다.

### 1. 인공신경망(Artificial Neural Network)

#### 가. 인공신경망 개요

인공신경망은 기존의 컴퓨터 구조로부터 과감하게 벗어나 생물학적 두뇌의 작동 원리를 그대로 모방하여 새로운 형태의 계산도구를 만들고자 하는 것이다(이재규 외, 1996). 즉 인간이 경험으로부터 학습해 가는 두뇌의 신경망 활동을 모방하여 자신이 가진 데이터로부터의 반복적인 학습 과정을 거쳐 패턴을 찾아내고, 이를 일반화함으로써 특히 향후를 예측하고자 하는 문제에 있어서 유용하게 사용되는 기법이다. 또한 인공신경망은 매우 복

잡한 구조를 지닌 데이터들 사이의 관계나 패턴을 찾아내는 유연한 비선형 모형의 하나로, 신경 생리학과의 유사성 때문에 일반적으로 다른 통계적 예측모형에 비해 흥미롭게 여겨지고 있다.

최초의 인공신경망 모형은 1957년 로젠블럿(Rosenblatt)에 의해 개발된 퍼셉트론(Perceptron)으로 퍼셉트론은 학습이 가능한 기계모형을 제시했고 여러 종류의 분류작업에 적용될 수 있다는 기대감에 많은 사람의 관심을 모았다(Rosenblatt, 1958). 1980년대 중반 입력층, 출력층 그리고 한 개 또는 그 이상의 은닉층을 구조로 하는 새로운 모형들이 제안되었으며, 특히 PDP(Parallel distributed processing)그룹에 의해 폭넓은 연구가 진행되었다. 이 그룹에서 제안한 모형은 은닉층과 백프로퍼게이션(back-propagation)학습알고리즘을 사용함으로써 선형분리문제 뿐만 아니라 여러 가지 문제점들을 해결할 수 있는 계기를 마련하였다. 백프로퍼게이션 학습알고리즘은 오차를 정정하는 규칙으로써 입력에 대해 원하는 반응과 실제로 얻어진 것들에 대한 차이를 줄여나가는 것이다(Bigus, 1996).

축적된 자료를 이용하여 독립변수와 종속변수간의 결합관계를 추출하여 패턴인식, 분류, 예측 등의 기능을 수행하는 인공신경망 모형은 입력변수와 출력변수가 연속형이나 이산형인 경우 모두를 다룰 수 있고, 입력변수와 결과변수의 관계를 정의하기 어렵고 복잡한 데이터에 대해서도 좋은 결과를 낼 수 있다는 장점 때문에 다양한 산업분야의 다양한 문제에 적용되고 있으며, 또한 기존의 통계적 기법에 비해 그 예측력이 우수함이 많은 학자들에 의해 증명되고 있다(이건창 등, 1994; Barniv et al., 1997; Bortiz and Kennedy, 1995; Chung and Tam, 1992; Salchenberger et al., 1992; Tam and Kiang, 1992; Wilson and Sharda, 1994).

### 나. 인공신경망 구조

입력계층(Input layer), 출력계층(Output layer) 그리고 은닉계층(Hidden layer)으로 이루어져 있는 인공신경망은 뉴런을 모형화한 처리요소들을 기본 구성단위로 하고 있다.

처리요소는 여러 다른 처리요소들로부터 입력을 받아들여 $Y_i$ 라고 표기된 단 하나의 출력값을 생성한 후, 이를 연결된 처리요소들에게 전달한다. 즉 $j$번째 처리요소가 $i$번째 처리요소로부터 전달받은 입력값을 $x_i$ 라고 표기하는데 $X_i$ 는 $i$번째 처리요소의 출력값(Y)이다. 생물학적 뉴런들간의 정보전달에 있어 시냅스가 중요한 역할을 담당하고 있듯이, 처리요소들간의 연결강도를 반영하기 위해 인공신경망에서는 연결가중치 혹은 단순가중치를 사용하고 이를 $W_{ij}$ 로 표기한다. 각 처리요소들은 전달받은 입력값들과 연결가중치(Synaptic weight)를 사용하여 다음 <식1>과 같이 먼저 순 입력값($net_j$)를 계산한 후, 이를 <식 2>를 이용하여 출력값을 결정한다.

$$net_j = \sum W_{ij} X_i \qquad <식1>$$

$$Y_j = f_j(net_j) \qquad <식2>$$

이때 순 입력값을 출력값으로 변환시키는 함수 $f_j$를 전이함수(Transfer function) 또는 활성함수(Activation function)라고 부른다. 인공신경망의 설계 시 전이함수의 결정은 매우 중요하며, 최근 인공신경망의 전이함수로 비선형함수들이 많이 사용되고 있고, 특히 시그모이드(Sigmoid)함수가 많이 사용되고 있다.

### 다. 인공신경망 특성

인공신경망은 다른 기법들과 비교하여 다음과 같은 특징들을 갖는다(이재규 외, 1996).

첫째, 인공신경망은 초 병렬 분산처리시스템으로서 정보의 저장, 처리 및 전달을 신경회로망 내의 특정부위들이 분담하는 것이 아니라 신경회로망 전체로 그 기능을 수행한다.

둘째, 인공신경망은 자기 조직화시스템(self-organization system)이다. 즉 인공신경망은 새로운 이미지나 패턴 또는 사례가 주어졌을 경우 이를 기억하기 위해 자동적으로 자신의 내부 상태를 바꾸지만 이용자의 입장에서 볼 때 인공신경망의 내부상태는 블랙박스(Black box)로 존재한다.

셋째, 인공신경망에서 학습된 지식은 뉴런(neuron)들간의 연결강도(connection strength)에 의해 분산 저장된다. 학습은 인공신경망에 있어 가장 중요하고도 기본적인 요소로서 인공신경망을 이용자가 원하는 상황에 적응하도록 훈련시킬 수 있음을 의미한다.

넷째, 인공신경망은 이용자의 측면에서 볼 때 기존의 컴퓨터처럼 복잡한 프로그래밍 과정을 요구하지 않는다. 이용자는 인공신경망의 구조 및 학습방법을 지정함으로써 간단히 원하는 결과를 얻을 수 있으며 인공신경망 내부의 처리절차도 매우 간단하게 이루어진다.

마지막으로 인공신경망은 패턴의 분류나 패턴 인식에 유용하며, 우리가 부딪치는 많은 문제들은 대부분 패턴 분류문제로 볼 수 있기 때문에 인공신경망의 응용범위가 매우 넓다고 볼 수 있다.

### 라. 인공신경망의 응용

인공신경망이 기존의 인공지능 및 전통적인 컴퓨팅과 다른 점은 전문가 시스템에서는 지식이 명확한 형태의 규칙으로 표현되어야 하지만 인공신경망은 예제를 통해 학습함으로써 자신의 규칙을 스스로 생성한다는 점이다. 인공신경망은 순차처리(sequential processing)에서의 병목현상, 현 인공지능 접근방식의 한계, 규칙표현과 메모리의 한계 등으로 기존의 접근방식에서는 어려웠던 부분을 표현하는데 좀더 유용하게 사용되어 질 수 있

어 최근에 많이 사용되고 있는 기법중의 하나로 그 사용범위가 넓어지고 있다.

즉, 축적된 자료를 이용하여 독립변수와 종속변수간의 결합관계를 추출하여 패턴인식, 분류, 예측 등의 기능을 수행하는 인공신경망 모형은 입력변수와 결과변수가 연속성이나 이산형인 경우 모두를 다룰 수 있고, 입력변수와 결과변수와의 관계를 정의하기 어렵고 복잡한 데이터에 대해서도 좋은 결과를 낼 수 있다는 장점 때문에 다양한 산업분야의 다양한 문제에 적용되고 있으며, 또한 기존의 통계적인 기법에 비해 그 예측력이 우수함이 많은 학자들에 의해 증명되고 있다(이건창 등, 1994; Barniv et al., 1997; Bortiz and Kennedy, 1995; Chung and Tam, 1992; Salchenberger et al., 1992; Tam and Kiang, 1992; Fletcher and Goss, 1993; Wilson and Sharda, 1994; Etheridge and Sriram, 1997).

## 2. 의사결정나무 분석(Decision Tree)

### 가. 의사결정나무 분석기법 개요

의사결정나무는 의사결정규칙(decision rule)을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석방법이다. 의사결정나무는 분석과정이 나무구조에 의해서 표현되기 때문에, 분류 또는 예측을 목적으로 하는 다른 방법들, 예를 들어, 신경망(neural network), 판별분석(discriminant analysis), 회귀분석(regression analysis) 등에 비해 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다(최종후 외, 1998).

의사결정나무 분석의 대표적인 알고리즘으로는 CHAID(Kass, 1980), CART(Breiman et al., 1984), C4.5(Quinlan, 1996) 등이 있으며, 이들은 SPSS, SAS 등 많은 소프트웨어 회사들에 의해서 다양한 제품으로 상용화 되어 있다.

### 나. 의사결정나무 모형의 구축

의사결정나무는 하나의 나무구조를 이루고 있으며, 마디(node)라고 불리는 구성요소들로 이루어져 있다. 의사결정나무는 뿌리마디로부터 시작하여 각 가지가 끝마디에 이를 때까지 자식마디를 계속적으로 분리, 형성해 나감으로써 완성된다.

(1) 분리기준

분리기준은 하나의 부모마디로부터 자식마디들이 형성될 때, 예측변수의 선택 시 범주의 병합이 이루어지는 기준을 의미한다. 즉, 어떤 예측변수를 이용하여 어떻게 분리하는 것이 목표변수의 분포를 가장 잘 구별해 주는지를 파악하여 자식마디가 형성되는데, 목

표변수의 분포를 잘 구별하는 정도를 순수도(purity) 또는 다른 분리기준에 의해서 측정한다.

(2) 정지규칙

정지규칙은 더 이상의 분리가 일어나지 않고 현재의 마디가 끝마디가 되도록 하는 여러 가지 규칙을 의미한다.

(3) 가지치기

지나치게 많은 마디를 가지는 의사결정나무는 새로운 자료에 적용할 때 예측오차(prediction error)가 매우 클 가능성이 있다. 따라서 형성된 의사결정나무에서 적절하지 않은 마디를 제거하여, 적당한 크기의 부나무(subtree)구조를 가지는 의사결정나무를 최종적인 예측모형으로 선택하는 것이 바람직하다. CART나 QUEST 또는 C4.5와 같은 알고리즘에서는 가지치기를 의사결정나무의 형성과정에 포함시키기도 하지만, 실제로는 연구자가 적절한 가지치기를 수행해 주는 것도 필요하다.

(4) 변수의 척도

의사결정나무를 형성할 때 변수가 가지고 있는 척도(measurement level)에 따라서 자료의 분리 또는 병합에 영향을 받는다. 또한 목표변수나 예측변수들이 어떤 척도를 가지고 있느냐에 따라서 사용할 수 있는 분석 알고리즘도 달라진다. 따라서 사용자는 각 변수의 특성을 정확히 파악하여 적절한 척도를 지정해 주어야 한다. 그 종류로는 명목형, 순서형, 연속형 등이 있다.

## Ⅳ. 실험설계

본 연구에서는 전술한 데이터마이닝 기법인 인공신경망과 의사결정나무 분석을 적용하여 스팸 메일을 분류해 내는 필터링 모형을 구축한다.

### 1. 표본데이터와 사용변수

모형의 실험을 위하여 4,601개의 이-메일 데이터를 활용하였다. 이 중 60.6%인 2,788개의 케이스는 비스팸(Non-Spam)메일로, 나머지 39.4%인 1,813개의 케이스는 스팸(Spam)메일로 구성하였다.

모형구축을 위하여 다음과 같은 과정을 통해 변수를 생성하였다. 우선 스팸 메일 판단에 도움이 될 것으로 보이는 단어들로 57개의 입력변수 후보군을 선정하고, 해당 단어나 문자가 메일의 문장 내에서 발견된 확률값을 축적된 데이터를 통하여 산정하였다. 모형에 사용될 최종 입력변수의 선정을 위하여 사용한 방법론은 통계분석을 통한 방법과 전문가로 하여금 선정하게 하는 방법이었다. 출력변수로는 스팸 메일 여부를 나타내는 1개의 변수를 사용하였다.

## 2. 입력변수의 선정

필터링 모형 구축을 위한 입력변수 선정을 위해 후보변수인 57개의 입력변수들 각각과 종속변수와의 단일변량 t-test를 실시하였다. t-test 결과 p-value 1% 미만인 변수 20개를 1차로 선정한 후 다시 다변량 판별분석을 위한 선택적 변수선정기법(step-wise selection)을 적용하여 최종적으로 9개의 변수를 선정하였다. 통계적 분석을 통해 선정된 변수들에 대한 설명과 특성은 다음의 <표 1>과 같다.

<표 1> 전문가집단 입력변수 선정결과표

| 변수명 | 변수설명 |
|---|---|
| E1 | '3d'가 나오는 확률 |
| E2 | 'remove'가 나오는 확률 |
| E3 | 'addresses'가 나오는 확률 |
| E4 | '000'이 나오는 확률 |
| E5 | 'technology'가 나오는 확률 |
| E6 | 'meeting'이 나오는 확률 |
| E7 | 're'가 나오는 확률 |
| E8 | '!'가 나오는 확률 |
| E9 | '$'가 나오는 확률 |

전문가들과의 인터뷰를 통해서도 10개의 변수를 선정하였다. 변수선정에 참여한 전문가 그룹은 4명의 웹 관리자들과 5명의 사내메일 전담자들, 그리고 1명의 스팸 메일 연구자로 구성되었다. 선정된 변수들에 대한 설명과 특성은 다음의 <표 2>와 같다.

<표 2> 통계적 방법으로 선정된 변수

| 변수명 | 변수설명 |
|---|---|
| X1 | 'make'가 나오는 확률 |
| X2 | '3d'가 나오는 확률 |
| X3 | 'order'가 나오는 확률 |
| X4 | 'addresses'가 나오는 확률 |
| X5 | 'money'가 나오는 확률 |
| X6 | 'technology'가 나오는 확률 |
| X7 | 'original'이 나오는 확률 |
| X8 | '!'가 나오는 확률 |
| X9 | '$'가 나오는 확률 |
| X10 | 대문자길이의 총합 |

## 3. 인공신경망 모형 구축

인공신경망 모형은 복합분석에 의한 변수선정방법을 통해 선정된 9개의 입력변수군으로 모형을 구축하였다. 총 4,601개 데이터를 인공신경망의 학습을 위한 학습용 자료(60%, 총 2,761개)와 테스트용 자료(20%, 920개), 검증용(20%, 920개)로 나누어 실험을 실시하였다.

인공신경망에서 학습용 자료는 인공신경망 모형을 학습하는데 사용되고, 테스트용 자료는 모형의 과도한 수렴을 막고 학습종료를 위한 최적의 조건을 찾기 위해 사용되며, 검증용 자료는 학습을 통해 구축된 모델의 검증을 위해서 사용한다. 모형구축에는 인공신경망 모형 구축 전문 툴(tool)인 NeuroShell 2(release4.0)을 이용하였다.

본 연구에서 사용한 인공신경망 모형은 다층 퍼셉트론(mult-layer perception)과 역전파 학습(back-propagation)알고리즘으로 입력계층과 출력계층, 그리고 하나의 은닉계층을 가지는 3층 퍼셉트론(three layer perceptron)을 사용하였으며, 은닉층의 노드(node)수는 입력변수와 동수를 사용하였다. 본 연구에서 사용된 모형의 출력층은 1개의 노드로 구성되어 스팸(1), 논스팸(0)을 나타내며, 모형의 출력값은 [0,1]의 범위에 존재한다.

## 4. 의사결정나무 분석 모형 구축

의사결정나무 분석 모형에서도 총 4,601개의 데이터를 모형구축용 자료 (80%, 총 3681개)와 검증용(20%, 920개)로 나누어 실험을 실시하였다. 모형구축을 위하여 사용한 툴(tool)은 Clementine 5.2.1으로, 의사결정나무분석 방법은 C.5 방법을 사용하였다. C.5는 1986년 Quinlan에 의해서 개발된 것으로 CART와 유사하나 CART와 달리 다지분리가 가능한 알고리즘이다(김지현, 1999).

# V. 실험결과 및 분석

## 1. 인공신경망 모형의 결과

복합 변수선정방법을 통해 선정된 9개의 입력변수군(A)과 전문가에 의해 선정된 10개의 변수군(B)에 의해 각각의 인공신경망 모형을 구축하였다.

총 4,601개 데이터를 인공신경망의 학습을 위한 학습용 자료(60%, 총 2,761개)와 테스트용 자료(20%, 920개), 검증용(20%, 920개)로 나누어 실험을 실시하였다. 은닉층의 노드 수는 입력층의 노드 수와 동수(set1), 입력노드 수의 2배수(set2), 입력노드 수의 반수(set3)로 세 가지 조건에서 다르게 실시하였다.

실험결과에 의하면 동일한 자료 내에서 은닉층의 노드 수만 다르게 한 결과 학습용, 테스트용, 검증용 데이터에서 근소한 차이를 보였으나 입력변수 A군의 경우에는 은닉노드수를 2배로 한 경우가 적중률이 가장 높았고, 입력변수 B군의 경우에는 은닉층 노드수를 입력노드 수의 반수로 한 경우가 적중률이 가장 높았다. 또한 전문가가 선정한 입력변수로 실험한 결과보다는 복합적 통계방법에 의해 선정된 입력변수가 더 높은 적중률을 보였다(<표 3> 참조).

<표 3> 인공신경망 모형별 적중률

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| A군 | Set1 | 훈련용 | 85.95 % | B군 | Set1 | 훈련용 | 83.39 % |
| | | 테스트용 | 86.87 % | | | 테스트용 | 81.10 % |
| | | 검증용 | 86.93 % | | | 검증용 | 82.27 % |
| | Set2 | 훈련용 | 87.86 % | | Set2 | 훈련용 | 82.40 % |
| | | 테스트용 | 87.39 % | | | 테스트용 | 80.21 % |
| | | 검증용 | 88.31 % | | | 검증용 | 82.80 % |
| | Set3 | 훈련용 | 87.64 % | | Set3 | 훈련용 | 84.25 % |
| | | 테스트용 | 87.39 % | | | 테스트용 | 84.28 % |
| | | 검증용 | 85.36 % | | | 검증용 | 84.43 % |

## 2. 의사결정나무분석 모형의 결과

통계적으로 선정된 9개의 입력변수(A)로 실험한 결과 90.31% 의 적중률이 나왔고, 전문가집단이 선정한 10개의 입력변수(B)로 실험한 결과 86.38% 의 적중률이 나왔다. 변수군(A)를 이용하여 도출된 규칙의 예는 다음과 같다(<표 4> 참조).

<표 4> 의사결정나무분석 결과(rules)

Rules for 0:
Rule #1 for 0:
if X2 =< 0
and X9 =< 0.055
and X4 =< 0.25
and X8 =< 0.092
then -> 0

Rule #2 for 0:
if X2 =< 0
and X9 =< 0.007
and X4 =< 0.25
and X8 > 0.092
and X8 =< 0.378
then -> 0

Rule #3 for 0:
if X2 =< 0
and X9 =< 0.055
and X4 =< 0.25
and X8 > 0.378
and X8 =< 0.775
and X7 =< 0.58
and X3 =< 0.07
then -> 0

Rule #4 for 0:
if X2 =< 0
and X9 =< 0.055
and X4 =< 0.25
and X8 > 0.378
and X8 =< 0.792
and X7 > 0.58
then -> 0

Rules for 1:
Rule #1 for 1:
if X2 =< 0
and X9 > 0.007
and X9 =< 0.055
and X4 =< 0.25
and X8 > 0.092
and X8 =< 0.378
then -> 1

Rule #2 for 1:
if X2 =< 0
and X9 =< 0.055
and X4 =< 0.25
and X8 > 0.378
and X8 =< 0.775
and X7 =< 0.58
and X3 > 0.07
then -> 1

Rule #3 for 1:
if X2 =< 0
and X9 =< 0.055
and X4 =< 0.25
and X8 > 0.775
and X7 =< 0.58
then -> 1
Rule #4 for 1:
if X2 =< 0
and X9 =< 0.055
and X4 =< 0.25
and X8 > 0.792
and X8 =< 4.347
and X7 > 0.58
and X7 =< 2.77
and X6 =< 0.89
then -> 1

## 3. 기법간 결과 비교 분석

인공신경망과 의사결정나무 기법을 통해 구축된 모형의 성과를 비교해 보면, 인공신경망이 평균 84.94%, 의사결정나무가 평균88.34%로 의사결정나무가 보다 높은 예측력을 보였다. 알고리즘 특성상 인공신경망은 결과 도출 과정을 해석하는 것이 매우 어려우므로 스팸 메일을 분류하는 데 확실한 모형 설명과 유용한 변수들의 적절한 제시를 위해서는 분석과정과 결과에 대한 설명력이 뛰어난 의사결정나무분석이 보다 효율적이다.

일반적으로 현실세계에서는 비즈니스 영역별로, 예측하고자하는 목적에 따라 추출하고자 하는 지식의 유형이 다양하고 이에 따라 지식을 추출하는 방법도 다양하게 적용될 수

있다. 각각의 목적과 문제 해결에 적합한 데이터마이닝 기법을 사용하는 것이 유용할 것이다.

## Ⅵ. 결론

최근 스팸 메일로 인한 폐해로 말미암아 스팸 메일을 차단하고자 하는 많은 노력들이 경주되어 왔다. 기존의 스팸 메일 방지 솔루션은 이-메일주소나 IP주소를 입력하여 스팸을 차단하거나. 릴레이를 불허하는 형태로 되어있다. 그러나 수많은 이-메일 계정을 가진 스패머와 고정IP보다는 유동적인 가상 IP를 많이 사용하는 현실에서는 최적의 대응이 어려운 실정이다(http://www.sendmail.com).

본 연구에서는 데이터마이닝 기법을 이용하여 스팸 필터링 모형을 구축하였다. 분류예측에 유용한 기법들인 인공신경망과 의사결정나무분석의 기법을 적용해 본 결과, 인공신경망이 평균 84.94%, 의사결정나무가 평균88.34%로 둘 다 우수한 예측력을 보여 주었다. 특히 의사결정분석은 얻어진 결과에 대한 계층구조를 보여줌으로써 결과에 대한 설명력에 있어서 우수하였다.

그러나 본 연구는 다음과 같은 한계를 지닌다. 첫째, 본 연구에서 사용된 데이터가 영문 이-메일 데이터라는 점으로, 향후 국문 메일에도 적용 가능한 모형의 구축이 필요하리라고 본다. 둘째, 본 연구에서 제시된 모형의 적중율이 100%가 아닌 이상, 모형 적용 시 중요한 메일을 스팸으로 오인할 가능성이 존재한다는 점이다. 스팸 필터링 모형과 같이 1종 에러와 2종 에러로 인한 비용이 크게 상이한 경우, 스팸 분류 지점의 조정을 통하여 오분류 비용(mis-classification cost)을 최소화 시켜야 한다. 이 문제는 본 모형을 적용하는 조직 특성과 오분류 비용 최소화를 위한 최적화와 관련된 연구를 통하여 해결해야 할 것이다.

## 참고문헌

김지현(1999), 데이터 마이닝의 의사결정나무분석을 이용한 사례분석, 이화여대자대학교 대학원 석사학위논문.

이건창, 김명종, 김혁(1994), 기업도산예측을 위한 귀납적 학습지원 인공신경망 접근방법: MDA, 귀납적 학습방법, 인공신경망 모형과의 성과비교, 경영학연구, 23, 2, 109-144.

이재규, 최형림, 김현수, 서민수, 주석진, 지원철(1996), 「전문가시스템 원리와 개발」, 서울: 법영사.

정재윤 (2000), 「이메일 마케팅」, 서울: 웹마니아.

최종후, 한상태, 강현철, 김은석(1998), 「데이터 마이닝 의사결정나무 분석」, 서울: SPSS 아카데미.

Barniv, R., Agawal, A. and Leach, R.(1997), Predicting the outcome following bankruptcy filing: A three-state classification using neural networks, *Intelligent Systems in Accounting , Finance and Management*, 6, 177-194.

Bigus, J. P., *Data Mining with Neural Networks*, Computing McGraw-Hill, 1996.

Bortiz, J. and Kennedy, D.(1995), Effectiveness of neural netwoks types for prediction of business failure, *Expert Systems with Applications*, 9, 503-512.

Breiman, l., Friedman, J. H., Olshen, R. A. and Stone, C. J.(1984), Classification and regression trees, Wadsworth, Belmont.

Chung, H. and Tam, K.(1992), A comparative analysis of inductive learning algorithm, *Intelligent Systems in Accounting, Finance and Management*, 2, 3-18.

Cranor, L. F. and LaMacchia, B. A.(1998) Spam!, *Communications of the ACM*, 41, 8, 74-83.

Etheridge, H. and Sriram, R.(1997), A comparative of the relative costs of financial distress models: Artificial neural networks, logit and multivariate discriminant analysis, *Intellgent Systems in Accounting, Finance and Management*, 6, 235-248.

Fletcher, D. and Goss, E.(1993), Forecasting with neural networks: An application using bankruptcy data, *Information and Management*, 24, 3, 159-167.

Gillin, P.(1999), Spam Spam, *computerworld*, 31, 22, 84-97.

Kass, G.(1980), An exploratory technique for investigation large quantotoes of categorical data, *Applied Statistics*. 29, 2, 119-127.

Quinlan, J. R.(1996), Learning Decision Tree Classfiers, *ACM Computing Surveys*, 28, 1.

Rosenblatt, F.(1958), The perceptron: Aprobabilistic model for information storage and organization in the brain, *Psychological Revenue*, 65, 386-408.

Salchenberger, L., Cinar, E. and Lash, N.(1992), Neural networks: A new tool for predicting thrift failures, *Decision Sciences*, 23, 899-916.

Tam, K. and Kiang, M.(1992), Managerial applications of neural networks: the case of bank failure predictions, *Management Science*, 38, 7, 926-947.

Wilson, R. and Sharda, R.(1994), Bankruptcy prediction using neural networks, *Decision Support Systems*, 11, 5, 545-557.

경영과컴퓨터(1999. 7), 사이버공간의 오염, 서울: 경영과컴퓨터.

한국경제(1999. 7. 5), 스팸 메일 법제재 나섰다, 서울: 한국경제.

http://www.cauce.org.

http://www.panix.com.

http://www.sendmail.com.

<Abstract>

# Spam Mail Classification Modeling Using Data Mining Techniques

Kyung-Shik Shin* · Su-San Ahn**

Due to the proliferation of unsolicited bulk e-mail, commonly referred to as "spam," ISPs have been deluged with complaints. This paper proposes the data mining approaches such as artificial neural networks and decision trees to build a spam mail classification model. Our preliminary results show that this approach is promising.

**Key words** : Data Mining, Artificial Neural Networks, Decision Tree, Spam Mail, Classification, Prediction

* Professor, College of Business Administration, Ewha Womans University
** College of Business Administration, Ewha Womans University