응용통계연구 제 13권 2호, 2000년, pp. 489-503

데이터 마이닝에서의 폴리클라스

구자용[1] 박헌진[2] 최대우[3]

요 약

다양한 형태의 데이터로부터 의사 결정에 유용한 정보 및 지식을 발견하려는 일련의 데이터분석 및 모형 선정과정을 데이터 마이닝(Data Mining)이라고 할 수 있다. 데이터 마이닝의 적용 예로는 신규고객에 대한 신용평가, 고객이탈방지 등과 같은 분야에서 발생하는 스코링 문제를 들 수 있는데 신용평가에서는 신용이 나쁠 가능성을 스코어로 나타내고 스코어가 높은 고객을 대상으로 특별관리를 할 수 있을 것이며 고객이탈방지에서는 이탈가능성을 스코어로 나타내고 스코어가 높은 고객을 대상으로 이탈 방지 캠페인을 벌일 수 있을 것이다. 본 논문에서는 스코링 문제를 사후확률에 대한 모형화 문제로 파악하였다. 폴리클라스를 스코링 문제에 적용하는 방법을 소개한 후 이를 독일 신용데이터, 국내 모 PC 통신회사 데이터 및 국내 모 이동통신 데이터에 적용하였다. 스코링의 성능은 이득률을 이용하여 평가하고자 하는데 나무 모형에 비하여 폴리클라스 방법이 우수함을 확인하였다.

주요용어: 다차원의 저주, 스코링, 이득률, 텐서 스플라인, 함수추정.

1. 머리말

현대 사회의 복잡성으로 우리는 엄청난 양의 정보를 접하며 살고 있으며 컴퓨터의 발전과 더불어 이러한 정보는 데이터 베이스로 구축되어 주어지게 된다. 이러한 방대하고 다양한 형태의 데이터로부터 의사 결정에 유용한 정보 및 지식을 발견하려는 일련의 데이터 분석 및 모형 선정과정을 데이터 마이닝(Data Mining)이라고 한다. 데이터 마이닝을 사용한 사례로는 보험요율산정, 개인신용평가, 신용카드 부정거래자 색출, 데이터 베이스 마케팅, 텔리커뮤니케이션 서비스 등을 들 수 있다. 이러한 사례를 통하여 볼 때 저자들은 데이터 마이닝 과정이 크게 계획(Design), 탐색(Exploration), 표현(Layout), 처리(Process) 및 분석(Analysis)으로 이루어진다고 생각한다 (그림 1.1 참조). 즉 데이터 마이닝에서는 계획단계에서는 문제 제기를 하고, 탐색단계에서는 데이터의 특성을 찾게 되며 표현에서는 여러가지 형태의 컴퓨터 그래픽에 의한 특성를 표현하고자 하며 이를 처리하고 분석하는 단계를 거치게 된다 (구자용, 박헌진, 최대우 1999). 여기서 데이터 마이닝 작업을 수행함에

1) (200-702) 강원도 춘천시 옥천동 1번지, 한림대학교 정보통계학과, 교수
E-mail: jykoo@sun.hallym.ac.kr
구자용의 연구는 1999년도 한림대학교 학술연구조성비에 의하여 이루어졌음.
2) (402-751) 인천시 남구 용현동 253 번지, 인하대학교 통계학과, 부교수
E-mail: hjpark@anova.inha.ac.kr
3) (449-791) 경기도 용인시 모현면 왕산리 산89, 한국외국어대학교 정보통계학과, 조교수
E-mail: dachoi@dreamwiz.com

있어서 사용되는 데이터 마이닝 도구는 특히 고성능의 그래픽 기능을 갖추고 있어야 하며 여러가지 상황에 적용가능하기 위하여 유연성이 높아야 한다. 여기서 유연성이란 다양한 함수형태에 적합이 가능한 정도를 의미하는데 (Stone 1985), 예컨데 직선은 이러한 의미로 유연성이 떨어진다고 할 수 있다.

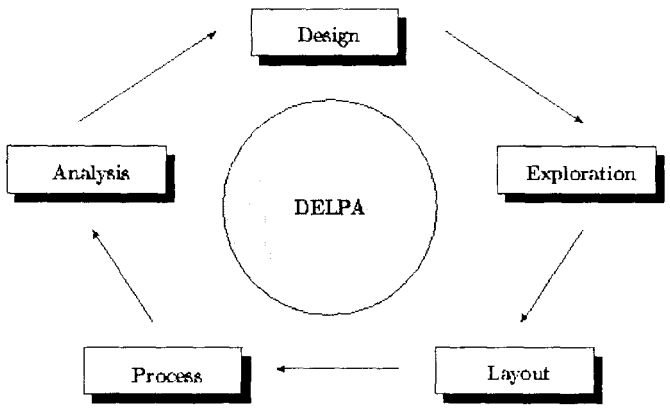

그림 1.1: 데이터 마이닝 단계

데이터 마이닝에서 통계적 방법을 적용할 때 데이터 마이닝 작업에서 고려할 상황의 다양성으로 전통적인 모수적 방법을 적용하기는 어려움이 많으므로 최근 컴퓨터의 발전과 더불어 활발히 연구되고 있는 비모수적 방법을 사용하는 것이 대부분이다. 모수적 방법이란 데이터를 생성하는 확률모형을 몇개의 모수로 모형화하는 방식이며 비모수적 방법에서는 이 확률모형을 무한개의 모수 즉 함수로 적합하는 방식를 일컫는데 전통적인 분포무관(distribution free) 방법 또는 순서방법(rank method)과는 구별된다. 이러한 특성으로 비모수적 방법이라는 말보다는 함수추정이라는 용어가 더 정확한 표현이라고 판단한다.

현재 데이터 마이닝에 쓰이고 있는 통계적 방법론 중의 하나는 Stone이 주도한 스플라인 방법론이라고 할 수 있다. 통계학의 함수추정 분야에서 이론적으로, 실제로 중요한 과업중의 하나는 일반화선형모형(Generalized Linear Models: GLMs)를 확장하는 것인데 GLMs의 주요 특성(정규 회귀, 로지스틱 회귀, 포아송 회귀 포함)을 고려하여 반응변수와 명목형 및 연속형 공변량을 포함하는 다변량 데이터를 다룰 수 있는 방향으로 전개되어 왔다. 여기서 GLMs에 대한 참고도서로는 McCullugh & Nelder (1989)가 있다. Stone은 이러한 방향의 연구를 수행하는 데에 있어서 주로 다항 스플라인 및 그들의 텐서 곱을 이용하였는데 Wahba의 평활 스플라인 접근방식과 구별하기 위하여 회귀 스플라인 방식이라는 패러다임을 제시하였다.

회귀 스플라인 패러다임에 속하는 초기연구로는 Agarwal & Studden (1980), Smith (1982), Koo (1990, 1992), Koo & Lee (1994) 등이 있다. 이러한 논문들에서 연구한 추정량들은 공변량의 수가 커지면 **다차원의 저주** (curse of dimensionality)를 피할 수 없게 된다. Hastie와 Tibshirani는 이러한 문제로 주효과만을 추정하고자 일반화 가법모형(Generalized

Additive Models: GAMs)을 제안하였는데 이에 대한 참고도서로는 Hastie & Tibshirani (1990)이 있다. Stone & Koo (1986), Friedman & Silverman (1989), Breiman (1993) Stone (1985, 1986) 등은 GAMs에 대해 가법 스플라인(additive spline) 추정량의 특성에 대한 연구를 수행하였다. 다차원의 저주를 피하면서 주효과만을 확인할 수 밖에 없는 GAMs의 한계를 극복할 수 있는 방법으로 제안한 방법이 Friedman (1991)의 MARS(Multivariate Adaptive Regression Splines)이며 이에 대한 이론 연구로는 Stone(1994), Stone *et al* (1997), Huang (1998) 등이 있다. 그외의 회귀 스플라인 패러다임에 따르는 연구로는 Kooperberg, Stone & Truong (1995), Kooperberg, Bose & Stone (1997), Koo (1997), Koo & Lee (1998) 등이 있다.

본 논문에서는 회귀 스플라인 방법 중의 하나인 폴리클라스 (Kooperberg, Bose & Stone 1997) 방법을 사용한 스코링(scoring) 문제에 대하여 연구하고자 한다. 데이터 마이닝의 적용 예로는 신규고객에 대한 신용평가와 같은 분야에서 발생하는 스코링 문제를 들 수 있는데, 예컨데 신용이 나쁠 가능성을 스코어로 나타내고 스코어가 높은 고객을 대상으로 특별관리를 할 수 있을 것이다. 이러한 상황에 적용가능한 데이터 마이닝 기법으로 폴리클라스 방법을 사용하고자 한다. 폴리클라스는 기존의 로지스틱 회귀방법을 확장한 것으로 선형 스플라인과 이들의 텐서 곱을 사용하여 다항 반응변수를 포함하는 데이터를 설명할 수 있는 방법이다. 폴리클라스는 사후확률을 직접 모형화하므로 최근 데이터 마이닝에서 많이 사용되고 있는 나무 모형과 같은 판별방식에 비해 사후확률이 필요한 경우에 특히 유용한 판별방법이다.

논문 구성은 다음과 같다. 제 2장에서는 데이터 마이닝의 개념을 주로 통계학적 관점에서 정립하고자 한다. 폴리클라스 방법은 제 3장에서 간략히 소개하고자 한다. 제 4장에서는 스코링을 정의하고 스코링 방법들의 성능은 이득률을 이용하여 평가하고자 한다. 여기서는 실제 데이터의 분석을 통하여 폴리클라스에 의한 스코링 방법이 나무 모형에 의한 방법보다 성능면에서 우월함을 규명하고자 한다.

2. 데이타 마이닝이란

데이터 마이닝은 컴퓨터 과학의 인공지능(artificial intelligence), 로봇비젼(robot vision), 패턴인식 등에 활용되는 기계학습(machine learning) 이론에서부터 시작되었다. 예를 들어 카메라를 통해 읽혀진 문자를 적당한 데이터의 형태로 변환시킨 뒤 과연 어떠한 문자가 입력되었는가를 판단하는 문자인식에서는 신경망을 사용하여 분류모형을 설정하고 있다. 신경망 등을 이용한 분류작업을 사람의 행위 및 이력데이터에 적용하여 모형을 세운 뒤 개인신용이나 해지여부 등을 미리 예측하는 작업을 데이터 마이닝이라 칭하기 시작한 것이다. 기계학습 이론에서는 신경망 이외에도 나무모형 등 데이터 마이닝에 적용될 수 있는 다양한 알고리즘들이 있다. 그러나 데이터 마이닝을 한마디로 데이터분석 및 예측모형 적합이라고 할 수 있으므로 기존의 통계학의 틀에서 크게 벗어난 것이 없다고 할 수 있다. 그리고 데이터 마이닝에서 활용되는 모형들은 이미 통계학의 유연(flexible) 함수 추정 분야에서 다루고 있는 내용들이다. 예컨데, 투영 탐색 회귀(Projection Pursuit Regression),

CART(Classification And Regression Tree), GAMs, MARS 등이 이에 속한다.

데이터 마이닝은 통계학은 물론 컴퓨터 과학, 경영정보학 등 여러 학문 분야에서 연구되고 있어 그 정의도 다양할 수 밖에 없다. 각분야에서의 정의를 종합하여 소개하면 다음과 같다. 데이터 마이닝이란 패턴 인식기술뿐 아니라 통계적·수학적 기법을 이용하여 저장된 거대한 데이터로부터 우리에게 유익하고 흥미있는 새로운 관계·성향·패턴 등 다양하고 가치있는 정보를 찾아내는 일련의 과정이라고 생각한다. 한 걸음 더 나아가 데이터 마이닝은 발굴된 값진 정보를 사용자가 전문적 지식없이 사용할 수 있도록 자동적으로 제공하는 시스템 개발과정까지 포함하기도 한다.

데이터 마이닝에서 다루는 주제로는 분류, 군집, 연관(association), 예측 등이 있는데 결국 데이터의 요약, 군집화, 분류화, 관계화, 성향(trend), 패턴인식(pattern recognition) 등을 통해 우리가 원하는 정보의 형태를 어떠한 방법으로 얻을 것인가가 중요한 연구과제인 것이다. 보다 자세한 내용은 최대우, 박일용, 박헌진 (1998)에 설명되어 있다.

데이터 마이닝이 활용될 수 있는 대표적인 분야로는 CRM (Customer Relationship Management)을 들 수 있다. 최근 고객의 다양한 정보를 거대한 데이터베이스(database, 이하 DB)로 보유하고 있는 백화점, 신용카드 회사, 이동 통신회사, 자동차 보험사 등에서는 DB로부터 목표 고객(target customer)의 특성을 파악하고 밝혀진 개개인의 특성에 적합한 마케팅 전략을 구사하고 있다. CRM을 구현하는 방법에는 여러가지가 있으나, 통계적 모형을 사용하여 목표 고객을 선별하고 그 특성을 파악하는 것이 가장 효과적이다. 즉, 엄청난 정보로 부터 고객의 행동패턴을 분석하고 유용한 정보를 발굴하는 분석능력이 CRM 성공여부의 핵심인 것이다. 이와 같이 거대 데이터로부터 주어진 데이터를 탐색(exploration)하고 특성별로 분할(segmentation)하고 예측(prediction)하기 위한 모형을 도출하는 등의 모든 데이터분석 행위가 데이터 마이닝인 것이다. 보험 요율산정에는 물론 개인 신용평가, 신용카드의 부정거래자 색출, 무선통신의 망관리 등 다양한 분야에 데이터 마이닝을 적용하여 큰 성과를 얻은 사례를 어렵지 않게 찾아볼 수 있다.

사실 데이터분석을 위한 여러 통계분석 기법 중 데이터 마이닝만을 위한 분석방법이 개발되어 있는 것은 아니다. 그리고 기존의 회귀분석이나 시계열 분석의 ARIMA와 같은 모수적 방법이 데이터 마이닝 분석에서 활용될 수 없다는 것도 아니다. 그러나, 데이터 마이닝이라는 작업이 주로 대용량의 데이터 및 복잡한 형태의 데이터를 대상으로 분석이 이루어지기 때문에 그에 적절한 통계분석 기법을 사용해야 한다.

대부분의 고전적 통계분석 모형에서 기본 가정을 하는 이유는 정보의 손실을 감수하면서도 분석결과에 대한 해석을 용이하게 하기 위해서이다. 그러나 데이터 마이닝에서는 결과에 해석보다 정확한 결과 예측이 중요할 수 있으므로 통계적 가정이 완화된 함수추정 방법을 주로 사용한다. 한편 반응변수와 설명변수의 관계를 밝히고 해석하는 것이 주된 목적인 데이터 마이닝에서는 고전적인 통계분석방법을 사용하거나 나무모형을 통해 규칙을 밝혀야 할 것이다.

3. 폴리클라스의 소개

본 절에서는 폴리클라스 방법을 소개하고자 하는데 보다 자세한 내용은 Kooperberg, Bose & Stone (1997)에 있으며 컴퓨팅 시간 및 타 방법과의 성능비교에 대한 사항은 Lim & Loh (1998)를 참조할 수 있다.

3.1. 폴리클라스 모형

폴리클라스 모형에는 K개의 수준을 갖고 있는 명목형 반응변수 Y와 이를 설명하는 데 쓰이는 예측변수 $x_1, \ldots, x_M$가 있다. 반응변수 Y가 택하는 값의 집합은 $\mathcal{K} = \{1, \ldots, K\}$로 나타낼 수 있으며 예측변수 $\boldsymbol{x} = (x_1, \ldots, x_M)$가 택하는 값의 집합은 $\mathbb{R}^M$의 부분집합인 $\mathcal{X}$로 나타낼 수 있다. 이때 예측변수가 확률변수 $\boldsymbol{X}$의 분포에 따른다고 하면 반응변수와 예측변수는 확률변수 쌍 $(\boldsymbol{X}, Y)$을 구성한다. $\boldsymbol{x} \in \mathcal{X}$이고 $k \in \mathcal{K}$에 대하여 조건부확률 $P(Y = k|\boldsymbol{X} = \boldsymbol{x})$가 양수값을 가진다고 가정할 때

$$\theta(k|\boldsymbol{x}) = \log \frac{P(Y = k|\boldsymbol{X} = \boldsymbol{x})}{P(Y = K|\boldsymbol{X} = \boldsymbol{x})}, \qquad x \in \mathcal{X}, \quad k \in \mathcal{K}$$

라 하자. 그러면 $x \in \mathcal{K}$에서 $\theta(K|\boldsymbol{x}) = 0$이며

$$P(Y = k|\boldsymbol{X} = \boldsymbol{x}) = \frac{\exp\theta(k|\boldsymbol{x})}{\exp\theta(1|\boldsymbol{x}) + \cdots + \exp\theta(K|\boldsymbol{x})}, \qquad \boldsymbol{x} \in \mathcal{X}, \quad k \in \mathcal{K} \tag{3.1}$$

와 같이 나타내어진다. 이때 식 (3.1)을 다항 회귀모형(polychotomous regression model)라 부르고자 하는데 $K = 2$이면 특별히 로지스틱 회귀모형이라고 하자.

보통의 로지스틱 회귀모형에서는

$$\theta(k|\boldsymbol{x}) = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kM}x_M, \quad 1 \le k \le K$$

형태의 선형, 가법모형을 사용한다. 그러나 예측변수의 효과를 비선형 함수로 모형화하는 것이 바람직할 수 있다 (Hastie & Tibshirani 1990). Koo & Lee (1994)은 텐서 스플라인을 이용하여 $M = 2$인 경우 반응함수를 추정하는 문제를 고려하였으며 Kooperberg, Bose & Stone (1997)에서는 GAMs를 더욱 일반화한 함수의 ANOVA형태로 분해를 고려하고 있는데 GAMs가 주효과만을 모형화할 수 있는데 반해 이러한 연구들은 교호작용도 모형화할 수 있는 장점이 있다. 폴리클라스에서는 기본적으로 함수들을 특정 기저함수로 전개하여 반응함수 $\theta(k|\boldsymbol{x})$를 추정하게 된다. 그러므로 추정하려는 함수 $\theta(k|\boldsymbol{x})$를 임의의 기저함수 $B_1, \ldots, B_p$로 전개하였을 때

$$\theta(k|\boldsymbol{x}) = \theta(k|\boldsymbol{x};\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_{jk}B_j(\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathcal{X}, \quad k \in \mathcal{K} \tag{3.2}$$

라 쓸 수 있다. 여기서 $1 \le k \le K-1$에 대하여 $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kp})^T$이고 $\boldsymbol{\beta}_K = 0$이다. 다시 말해서 $p(K-1)$차원의 열벡터를 $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K-1})$라고 정의하면

$$P(Y = k|\boldsymbol{X} = \boldsymbol{x};\boldsymbol{\beta}) = \frac{\exp\theta(k|\boldsymbol{x};\boldsymbol{\beta})}{\exp\theta(1|\boldsymbol{x};\boldsymbol{\beta}) + \cdots + \exp\theta(K|\boldsymbol{x};\boldsymbol{\beta})}, \qquad \boldsymbol{x} \in \mathcal{X}, \quad k \in \mathcal{K}$$

형태의 모형을 이용하여 반응변수와 예측변수들간의 관계를 모형화하고자 한다.

폴리클라스는 $\theta(k|\boldsymbol{x})$에 대한 기저함수로 스플라인 (spline) 함수와 그들의 2차 텐서곱(tensor product)을 사용한다. 이때 스플라인함수는 다음과 같이 정의된다.

$$(x_i - t_{ik})_+ = \begin{cases} x - t_k & (x_i \geq t_k\text{일 때}) \\ 0 & (\text{그외}) \end{cases}$$

이러한 기저함수와 2차 텐서곱을 위의 추가단계와 삭제단계를 수행하면서 각각의 기저함수에 대한 계수 $\boldsymbol{\beta}$는 최대우도추정법을 사용하여 $\hat{\boldsymbol{\beta}}$를 추정하고 Y에 대한 추정값으로 $\theta(k|\boldsymbol{x};\hat{\boldsymbol{\beta}})$를 최대로 하는 k값으로 선택한다. 이것은 단위비용(unit cost)을 갖을 경우 베이지안 다중판별법(Bayes multiple classification rule) 즉, 조건부확률 $P(Y = k|\boldsymbol{X})$를 최대로 하는 k값을 판별추정값으로 사용하는 것과 동일하다.

3.2. 최종모형 선택

앞의 (3.2)에서 먼저 고려되어야 할 문제는 어떤 종류의 기저함수를 사용하고 또 얼마나 많은 기저함수를 사용할지를 결정해야한다. 모형 G를 $\theta(k|\boldsymbol{x};\boldsymbol{\beta})$들의 집합이라고 하고 $\mathcal{G}$를 이들의 집합이라고 할 때

(a) $G \in \mathcal{G}$에서 공간 G의 차원은 $p \geq P_{min}$이며 최소차원을 갖는 공간 $G_{min} \in \mathcal{G}$은 유일하게 존재한다.

(b) 만약 $G \in \mathcal{G}$가 $p(> P_{min})$차원의 공간일때 G에 $p-1$차원 부분공간 $G_0 \in \mathcal{G}$가 적어도 하나는 존재한다.

(c) 만약 $G \in \mathcal{G}$가 $p(> P_{min})$차원의 공간일때 G를 포함하는 $p+1$차원의 $G_1 \in \mathcal{G}$가 적어도 하나는 존재한다.

우리는 조건(a)에 의해서 최소의 공간부터 모형을 적합하고 새로운 기저함수의 추가를 고려할 수 있다. 이것을 기저함수의 추가단계라 부르면 반대로 임의의 p차원에서 불필요한 기저함수를 삭제하는 것을 기저함수의 삭제단계라고 할 수 있다.

임의의 스플라인함수와 그 들의 2차 텐서곱으로 만들어지는 기저함수들의 수는 매우 많이 존재하게되어 임의의 p차원에서 $p+1$차원으로 공간 G를 확장할 때 고려해야할 변수가 너무 많아진다. 그러므로 가장 중요한 기저함수를 골라야만 한다. 폴리클라스에서는 다음과 같은 조건으로 추가될 기저함수를 제약한다.

(가) 항상 $x_i (i = 1, \ldots, p)$를 고려한다.

(나) 만약 x_i가 이미 모델에포함되어 있다면 $(x_i - t_{ik})_+$를 고려한다.

(다) 만약 x_i와 x_j가 이미 모델에 포함되어 있다면 그들의 텐서곱 $x_i x_j$를 고려한다.

(라) 만약 x_i,x_j와 $(x_j - t_{jk})_+$이 이미 모델에 포함되어 있다면 $x_i(x_j - t_{jk})_+$를 고려한다.

(마) 만약 $x_i(x_j - t_{jk})_+$와 $x_i(x_j - t_{jk})_+$이 모델에 포함되어 있다면 $(x_i - t_{ik})_+(x_j - t_{jk})_+$를 고려한다.

위의 (가)부터 (마)까지 조건은 예측변수 x_i가 먼저 모델에 포함되어야만 그 변수에 대한 스플라인 함수를 추가할 수 있으며 또한 기저함수의 텐서곱도 각각의 변수가 사전에 모델에 포함되어있을 때만 고려할 수 있다. 이러한 가정은 모델의 단순화와 함께 적합된 모델에 대한 해석을 용이하게 할 뿐아니라 기저함수 추가단계에서 발생되는 분산을 줄일 수 있게 된다.

폴리클라스에서 사용하는 최적의 모델을 선택하는 방법으로는 AIC, 검정(test) 데이터를 사용하는 방법 및 교차 타당성법(cross-validation)이 있는데 본 연구에서는 AIC를 이용한 방법을 고려하였다. 교차 타당성법은 특히 계산속도면에서 AIC에 비해 사용하기가 어려운 문제점이 있다. AIC를 이용한 방법은

$$AIC_{\alpha,\nu} = -2l_\nu + \alpha(K-1)p_\nu \tag{3.3}$$

과 같이 정의된 $AIC_{\alpha,\nu}$를 최소로 하는 $\hat{\nu}$를 구하고 이에 해당하는 모형을 최종모형으로 선택하는 방법이다. 여기서 α는 페널티 파라메터를 나타내며 l_ν는 p_ν개의 모수를 가지는, ν번째로 적합된 모델에 대한 로그가능성(log-likelihood)를 나타낸다.

4. 실제 데이터 분석 예들

본 장에서는 스코링 방법을 정의하고 마이닝 기법들에 의한 스코링 성능을 이득률에 의하여 비교하고자 한다. 데이터로는 독일 신용 데이터, 국내 모 PC 회사 데이터 및 국내 모 이동통신회사 데이터를 사용하였다.

4.1. 스코링 및 이득률

특정 데이터에서 미래의 데이터의 예측변수들의 값을 $\boldsymbol{x}_{\text{미래}}$라 하고 대응하는 반응변수의 값을 $Y_{\text{미래}}$라 하자. 이때 판별규칙을 세운다면 $P(Y_{\text{미래}} = 1|\boldsymbol{X} = \boldsymbol{x}_{\text{미래}})$에 대한 추정값을 구할 수 있고, 이 값이 높은 사람들을 감시 대상으로 하는 데이터 마이닝 솔루션을 제공할 수 있을 것이다. 이러한 문제에서 $P(Y_{\text{미래}} = 1|\boldsymbol{X} = \boldsymbol{x}_{\text{미래}})$에 대한 추정값을 신규 데이터의 스코어라고 정의하며 판별 규칙에 의해 스코어를 구하는 과정을 스코링이라고 정의할 때, 판별규칙이 데이터 마이닝 기법으로 의미를 가지려면 스코어에 대한 정밀한 추정값을 제시하여야 한다.

스코링에 대한 여러가지 방법이 있을 때 이들의 성능(performance)를 비교하기 위하여 이득률(gain)을 도입하고자 한다. 특정 판별규칙 D에 의하여 구한 사후확률 모형, 즉 $P(Y = 1|\boldsymbol{X} = \boldsymbol{x})$에 대한 추정 규칙을 $\hat{S}^D(\boldsymbol{x})$로 나타내고 기존 고객의 스코어를

$$s_i^D = \hat{S}^D(\boldsymbol{x}_i),\ 1 \le i \le n$$

라 하자. 이때 전체 데이터에서 실제로 관심의 대상이 되는 데이터 수를 N_1라 하고, s_i^D의 크기순으로 나열했을 때 상위 $p \times 100\%$ 데이터 중에서 실제로 관심이 되는 데이터들의 수

를 N_p^D라 하면 이득률은

$$\text{이득률}(p, D) = \frac{N_p^D}{N_1 \times p}, \qquad 0 < p < 1 \tag{4.1}$$

로 정의한다. 판별규칙 D가 아무런 판별력을 갖지 못하는 경우 이득률이 p에 상관없이 1에 가까와지며 반대로 판별력이 좋을 경우 그 값이 커지게 된다. 그러므로 이득률로 여러 판별규칙의 성능을 비교할 수 있다.

실제로 나무모형과 폴리클라스 방법를 비교하기 위한 스코링 알고리즘은 다음과 같다. 전체 N개의 데이터를 N_1개의 훈련데이터와 $N_2 = N - N_1$개의 테스트 데이터로 구분하여 각각의 모델을 훈련 데이터에 적합시킨 후 테스트 데이터를 이용하여 이득률을 구하게 된다. 여기서 훈련 데이터은 비복원 단순임의추출법에 의하여 구성하였다. 나무모형에 의한 이득률의 계산과정은 다음과 같다.

(a) 훈련(training) 데이터를 이용하여 나무를 생성한다.

(b) 디비언스(deviance) 그림을 보고 최적인 최종노드의 숫자를 결정한다.

(c) 테스트 데이터를 이용하여 이득률을 계산한다.

한편 폴리클라스에 의한 이득률의 계산과정은 다음과 같다.

(가) 훈련데이터를 이용하여 위에서 설명한 바와 같이 AIC를 이용하여 최적의 폴리클라스 모형을 결정한다.

(나) 테스트 데이터를 이용하여 이득률을 계산한다.

폴리클라스의 경우 모형 선택의 기준으로는 AIC를 사용하였으며 나무 모형은 가지치기(prunning)를 사용하지 않은 경우와 크기별 이탈도와 오분률 갯수에 근거한 가지치기를 사용한 경우를 고려하였다.

폴리클라스를 수행할 경우 명목형 설명변수는 S-Plus 패키지에 의하여 알파벳 순서로 자연수를 대응시킨후에 가변수(dummay variable)로 만든 후에 폴리클라스 알고리즘을 적용하는 방법이 일차로 고려할 수 있는 방법이다. 그런데 흥미로운 사실은 가변수로 변환하지 않고 명목형 변수를 연속형 변수로 간주하여 폴리클라스를 적용하는 것이 더 좋은 결과를 낳는 경우가 있다는 사실이다. 이러한 현상은 현재로서는 그 이유를 설명하기는 어려우나 명목변수의 수준(level)이 많을 경우 가변수를 사용하면 변수의 수가 늘어나서 폴리클라스와 같이 복잡한 모형을 적합할 때 다차원의 저주를 극복하기가 어렵다는 점에서 그 원인을 찾을 수 있다고 판단된다.

아래에서 각 그림에서 A는 명목형 변수를 연속형으로 간주하여 폴리클라스를 적용한 결과를, B는 명목형 변수를 가변수로 변환하여 폴리클라스를 적용한 결과를, C는 가지치기(pruning)를 한 나무모형의 이득률을, 그리고 D는 가지치기를 하지 않은 나무모형의 이득률을 각각 나타낸다. 한편, 이득률을 비교하는 영역은 p가 0.05에서 0.5 사이가 주된 관심이므로 이 부분에서 위의 4가지 방법을 비교하고자 한다.

4.2. 독일 신용 데이터

폴리클라스에 의한 신용평가 방법을 독일 신용 데이터에 적용함으로써 폴리클라스 스코링 방법을 설명하고자 한다. 독일 신용 데이터는 독일의 특정 신용카드회사에서 신규고객에 대한 신용도를 예측하기 위해 기존의 고객에 대한 여러가지 특성을 관측한 데이터이다. 고객신용을 나타내는 변수 Y는

$$Y = \begin{cases} 1 & (\text{신용이 나쁜 고객}) \\ 2 & (\text{신용이 좋은 고객}) \end{cases}$$

로 정의하고 예측변수로는 13개의 명목형 변수와 7개의 연속형 변수로 이루어져있다. 이 데이터에 대한 스코링 방법으로 웹상에서 구현된 S-Delpa를 사용하여 나무 모형과 폴리클라스 모형을 비교하고자 한다. 훈련데이터를 최종노드의 수는 4개가 최적인 것으로 판단된다. 독일 신용 데이터의 경우 $N = 1000$, $N_1 = 700$, $N_2 = 300$이다.

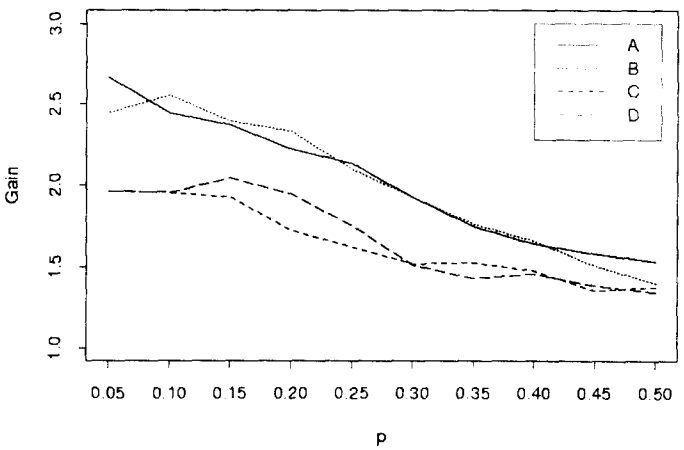

그림 4.1: 독일 신용 데이터에 대한 나무모형과 폴리클라스의 이득률비교.

그림 4.1에서는 독일 신용 데이터에 대해 나무모형에 의한 이득률과 폴리클라스에 의한 이득률을 보여준다. 폴리클라스에 의한 이득률이 전반적으로 나무모형에 의한 이득률보다 크므로 폴리클라스에 의한 스코링 방법이 월등히 좋다는 점을 확인할 수 있다. 폴리클라스를 실행할 때 명목형 변수의 경우 연속형 변수와 같이 처리하는 방법이 가변수(dummy variable)을 사용하는 방법보다 이득률 관점에서는 상대적으로 우월한 성능을 보임을 확인할 수 있다. 한편, 나무모형인 경우 가지치기를 한 결과가 이득률 관점에서 보면 우월한 결과를 보여줌을 확인할 수 있다.

4.3. 국내 모 PC 통신 회사 데이터

국내 모 PC 통신 회사는 서비스 해지자 예측을 위한 모델을 도출하기 위하여 과거 데이터를 이용하여 해지자 패턴을 분석하였다. 여기서 주된 관심은 이득률 관점에서 나무모형

과 폴리클라스의 성능을 비교하는 것이다. 훈련 데이터는 과금(billing) 데이터와 해지 여부 관찰 데이터로 구성되어 있다. 훈련 데이터는 98 년 10월부터 99년 4월 데이터로 이루어져 있으나 과금 데이터는 10월부터 1월까지의 4개월간 만을 사용하고 2, 3, 4월 3개월간의 해지 여부를 목표(target) 변수로 사용한다. 해지 여부를 나타내는 반응변수 Y는

$$Y = \begin{cases} 1 & (\text{서비스 지속 고객}) \\ 2 & (\text{서비스 해지 고객}) \end{cases}$$

로 정의하고 설명변수는 상품의 종류, 지역, 나이, 사용기간 등을 포함한 8개이다. 데이터 마이닝 작업에 사용된 데이터의 수는 총 $N = 56,580$ 건이고 그중 70%에 해당되는 $N_1 = 39,589$ 건이 훈련 데이터로 사용되었고 나머지는 모형 도출 후 이득률에 의한 모형의 성능 평가를 하기 위한 확인(validation) 데이터로 이용하였다.

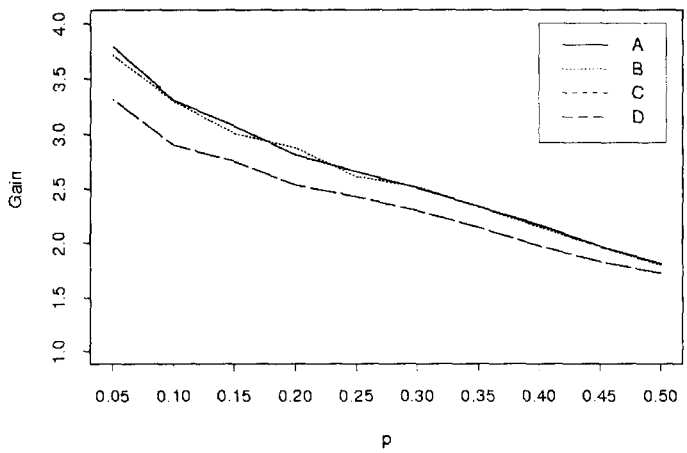

그림 4.2: 국내 모 PC 통신 회사 데이터에 대한 나무모형과 폴리클라스의 이득률비교.

그림 4.2은 국내 모 PC 통신회사 데이터에 대해 나무모형에 의한 이득률과 폴리클라스에 의한 이득률을 보여준다. 이 경우에도 폴리클라스에 의한 이득률이 전반적으로 나무모형에 의한 이득률보다 크므로 폴리클라스에 의한 스코링 방법이 월등히 좋다는 점을 확인할 수 있다. 독일 신용 데이터의 분석 결과와는 달리 폴리클라스를 실행할 때 명목형 변수의 경우 연속형 변수와 같이 처리하는 방법과 가변수를 사용하는 방법이 비슷한 결과를 나타냄을 발견할 수 있다. 한편, 나무모형인 경우 가지치기를 한 결과와 그렇지 않은 결과는 이득률 관점에서 보면 동일한 결과를 보여줌을 확인할 수 있다.

4.4. 국내 모 이동통신 회사 데이타

국내 모 이동통신 회사는 서비스 해지자 예측을 위한 모델을 도출하기 위하여 과거 데이터를 이용하여 역시 해지자 패턴을 분석하였다. 본 논문에서는 주로 이득률 관점에서 나무모형과 폴리클라스의 성능을 비교하고자 한다. 데이터 마이닝 작업에 사용된 데이터의

수는 총 $N = 32,258$ 건이고 그중 70%에 해당되는 $N_1 = 22,580$ 건이 훈련 데이터로 사용되었고 나머지는 모형 도출 후 이득률에 의한 모형의 성능 평가를 하기 위한 확인(validation) 데이터로 이용하였다. 설명변수로는 지역, 3개월간 평균 사용시간 등을 포함한 5개인데 이중 3개는 요인(factor) 변수이며 나머지지 2개는 연속형이며 반응변수는 해지 여부이다. 즉, 해지 여부를 나타내는 반응변수 Y는

$$Y = \begin{cases} 1 & (\text{서비스 지속 고객}) \\ 2 & (\text{서비스 해지 고객}) \end{cases}$$

로 정의된다.

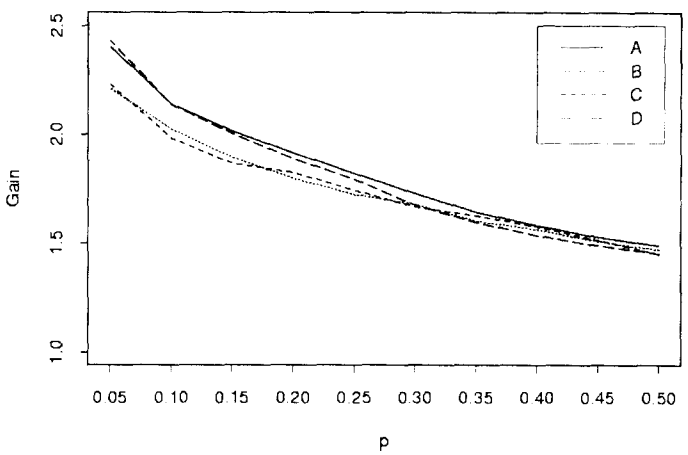

그림 4.3: 국내 모 이동 통신 회사 데이터에 대한 나무모형과 폴리클라스의 이득률비교 요인변수들도 포함한 경우.

그림 4.3은 국내 모 이동통신 회사의 데이터에 대해 나무모형에 의한 이득률과 폴리클라스에 의한 이득률을 보여준다. 이 경우에는 명목형 변수의 경우 연속형 변수와 같이 처리하는 폴리클라스에 의한 이득률이 가지치기를 시행한 나무모형의 이득률에 비하여 약간 우월하게 나타났다. 그런데 독일 신용 데이터 및 PC 통신회사 데이터의 분석 결과와는 달리 폴리클라스를 실행할 때 명목형 변수의 경우 가변수를 사용하는 방법이 이득률 관점에서 가지치기를 하지 않은 나무모형과 그 성능이 비슷하며 여타의 두 방법에 비하여서 성능이 나쁘게 나왔다. PC 통신회사 데이터에서 명목형 변수의 수준수는 8개를 넘지 않으나 이동통신 회사 데이터의 경우 수준수가 12개 내지 17개가 된다는 점이 가변수를 이용하는 방법의 성능이 떨어짐에 대한 한 이유가 되리라고 판단한다.

5. 맺음말

본 논문에서는 데이터 마이닝의 적용 예로써 신규고객에 대한 신용평가와 관련하여 발생하는 스코링 문제, PC 통신 서비스 해지와 관련한 스코링 문제 및 이동통신 서비스 해

지와 관련한 스코링 문제를 고려하였다. 데이터 마이닝 기법으로서 Kooperberg, Bose & Stone이 제안한 폴리클라스 방법의 장점을 나무모형과의 성능 비교로 확인하였다. 폴리클라스는 사후확률을 직접 모형화하므로 최근 데이터 마이닝에서 많이 사용되고 있는 나무모형에 의한 방법에 비해 본 논문에서와 같이 사후확률이 필요한 스코링 경우에 특히 유용한 판별방법임을 확인하였다. 또한, PC 통신 데이터와 이동통신 데이터의 경우 대용량의 데이터이면서 범주형 설명변수가 존재하는 다복잡한 형태의 데이터라고 할 수 있는데 본 논문에서 고려한 나무모형이나 폴리클라스 방법은 이러한 데이터에 적용이 가능함을 확인하였다.

본 논문에서 고려한 세가지의 데이터 분석 결과에 따르면 폴리클라스 방법이 이득률 관점에서 보면 나무모형에 의한 방법보다 우월한 성능을 보여줌을 확인하였다. 본 논문에서 그래프를 통하여 보여준 결과는 훈련데이터의 추출이 양호한 경우인데 그렇지 못한 경우에는 나무모형에 비해 일양적으로(uniformly) 우월한 성능을 보여주지 못하는 경우도 있었다. 물론 이 경우에도 나무모형보다 성능이 나쁜 영역은 p가 0.05이하인 경우 뿐이었다. 이와 같이 훈련 데이터 추출에 따르는 성능의 변동 문제는 bagging 방법(Breiman 1996)을 사용하면 추출에 따르는 변동성 문제를 극복할 것으로 기대되어 이에 대한 연구는 매우 의미있는 것으로 판단된다. 한편 추가적인 연구 주제로 중요하다고 판단되는 것은 명목형 변수의 처리 문제이다. 즉, 폴리클라스를 적용할 경우 가변수로 변환하는 것이 일차적으로 고려할 수 있는 방법이나 양의 정수로 대응한 후 연속형 변수로 처리하는 것이 우월한 성능을 보여주는 경우가 있다는 점이다. 이러한 현상이 신경망 방법을 포함한 비선형 방법들에서 나타나는 일반적인 것인가 하는 점과 그러한 현상에 대한 심층적인 연구는 매우 의미있는 연구라고 판단된다.

폴리클라스에 의한 스코링 방법은 다양한 분야에 적용할 수 있을 것이다. 먼저 신용평가와 관련해서는 신용이 나쁠 가능성을 스코어로 나타내고 스코어가 높은 고객을 대상으로 신용 관련하여 특별한 주의를 기울어야 할 것이다. 한편, 고객이탈방지에서는 특정서비스를 받기를 포기할 가능성을 스코어로 나타내고 스코어가 높은 고객을 대상으로 캠페인을 벌임으로써 고객 이탈을 최소화할 수 있을 것이다.

감사의 글

본 논문의 프로그래밍 작업을 도와 준 인하대학교 통계학과의 하동혁군과 한림대학교 정보통계학과의 황광연군에게 감사의 말을 전하고자 한다. 또한 본 논문에 대하여 여러 가지 도움 말씀을 주신 두 심사위원들에게도 감사의 말을 표하고자 한다.

참고문헌

[1] 구자용, 박헌진, 최대우 (1999). The development of data mining solution based on intranet. 정보과학의 통계학 응용 학술회의. 8-15.

[2] 최대우, 박일용, 박헌진 (1998). 데이터 마이닝을 이용한 자동차사고 다발자 성향분석. 보험개발연구. 보험개발원 보험연구소. 제3호. 1-24.

[3] Agarwal, G.G. and Studden, W.J. (1980). Asymptotic integrated mean square error using least squares and bias minimizing spline. *Ann. Statist.* **8** 1307-1325.

[4] Breiman, L. (1993). Fitting additive models to data. *Comput. Statist. Data Anal.* **15** 13-46.

[5] Breiman, L. (1996). Bagging predictors. *Machine Learning.* **26**, 123-140.

[6] Friedman, J.H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19** 1-141.

[7] Friedman, J.H. and Silverman, B.W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31** 3-39.

[8] Hastie T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

[9] Huang, J.Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *Ann. Statist.* **26**, 242-272.

[10] Koo, J.-Y. (1990). Optimal rates of convergence for tensor spline regression estimators, *J. Korean Statist. Soc.* **19** 105-112.

[11] Koo, J.-Y. (1992). Optimal rates of convergence in tensor Sobolev regression, *J. Korean Statist. Soc.* **21** 153-166.

[12] Koo, J.-Y. (1997). Spline estimation of discontinuous regression functions, *J. Comput. Graphical Statist.* **6** 266-284.

[13] Koo, J.-Y. and Lee, K.-W. (1998). B-spline estimation of regression functions with errors in variable. *Statist. Prob. Lett.* **40** 57-66.

[14] Koo, J.-Y. and Lee, Y. (1994). Bivariate B-splines in generalized linear models, *J. Statist. Comput. Simul.* **50** 119-129.

[15] Kooperberg, C., Bose, S. and Stone, C.J. (1997). Polychotonous regression. *J. Amer. Statist. Assoc.* **92** 117-127.

[16] Kooperberg, C., Stone, C.J. and Troung, Y.K. (1995). Hazard regression. *J. Amer. Statist. Assoc.* **90** 78-94.

[17] Lim, T.-S. and Loh, W.-Y. (1998). An empirical comparison of decision trees and other classification methods. Tehchnical report 979. Department of Statistics, University of Wisconsin, Madison.

[18] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

[19] Smith, P.L. (1982). Curve fitting and modeling with splines using statistical variable selection techniques. Report NASA 166034, Langley Research Center, Hampton, VA.

[20] Stone, C.J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689-705.

[21] Stone, C.J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590-606.

[22] Stone, C.J. (1994). The use of polynomial splines and their products in multivariate function estimation. *Ann. Statist.* **22** 118-184.

[23] Stone, C.J., Hansen, M., Kooperberg, C. and Troung, Y. (1997). Polynomial splines and their products in extended linear modeling (with discussion). *Ann. Statist.* **25** 1371-1470.

[24] Stone, C.J. and Koo, C.-Y. (1986). Additive spline in statistics. In *Proceedings of the Statistical Computing Section* 45-48. Amer. Statist. Assoc., Washington, DC.

[2000년 3월 접수, 2000년 7월 채택]

Polyclass in Data Mining

Ja-Yong Koo[1] Heon Jin Park[2] Daiwoo Choi[3]

ABSTRACT

Data mining means data analysis and model selection using various types of data in order to explore useful information and knowledge for making decisions. Examples of data mining include scoring for credit analysis of a new customer and scoring for churn management, where the customers with high scores are given special attention. In this paper, scoring is interpreted as a modeling process of the conditional probability and polyclass scoring method is described. German credit data, a PC communication company data and a mobile communication company data are used to compare the performance of polyclass scoring method with that of the scoring method based on a tree model.

Keywords: Curse of dimensionality; Function estimation; Gain; Scoring; Tensor-product spline.

1) Professor, Department of Statistics, Hallym University. E-mail: jykoo@sun.hallym.ac.kr
The research of Ja-Yong Koo was supported by a grant from Hallym University, Korea.
2) Associate Professor, Department of Statistics, Inha University. E-mail: hjpark@anova.inha.ac.kr
3) Assistant Professor, Department of Statistics, Hankook University of Foreign Stduies. E-mail: dachoi@dreamwiz.com