The Korean Communications
in Statistics Vol. 11 No. 2, 2004
pp. 255-264

A Study on Data Mining Using the Spline Basis

Sun Geun Lee[1], Songyong Sim[2], and Ja-Yong Koo[3]

Abstract

Due to a computerized data processing, there are many cases when we encounter a huge data set. On the other hand, advances in computing technologies make it possible to deal with a huge data set. One important area is the data mining. In this paper we consider data mining when the dependent variable is binary. The proposed method is to use the poly-class model when the independent variables consists of continuous and discrete variables. An example is provided.

Keywords : CRM, Data Mining, Spline, Gain, Polyclass

1. 서론

사회가 발전하고 각종 자료의 양이 기하급수적으로 늘어감에 따라 그에 따른 분석방법이 필요하게 되었다. 특히 컴퓨터의 발달로 인하여 기업 등 많은 부분에서 모여진 자료를 데이터베이스로 구축하여 관리 및 필요한 정보를 수집하게 되었다. 점차적인 데이터베이스의 다양화 및 거대화로 인하여 적시적소에 필요한 정보를 찾아내는 과정이 필요하게 되었고 이는 데이터마이닝이 생기게 된 이유 중의 하나이다.

데이터마이닝을 사용한 사례로는 보험효율 산정, 개인신용평가, 신용카드 부정거래자 색출, 데이터베이스 마케팅, 텔레커뮤니케이션, 서비스 등을 들 수 있다. 이러한 사례를 통하여 볼 때 데이터마이닝 과정은 크게 계획(Design), 탐색(Exploration), 표현(Layout), 처리(Process), 분석(Analysis)으로 이루어진다. 계획단계에서는 문제 제기를 하고, 탐색단계에서는 데이터의 특성을 찾게 되며 표현 단계에서는 여러 가지 형태의 컴퓨터 그래픽에 의한 특성을 표현하고자 하며 이를 처리하고 분석하는 단계를 거치게 된다(장남식(1999)).

데이터마이닝에서 전통적이며 가장 많이 알려진 통계적 방법으로는 회귀분석(regression analysis)이나 판별분석(discriminant analysis)이 있다. 이외에도 주로 다루는 방법으로는 분류, 군집, 연관, 예측 들이 있는데 이는 결국 데이터의 요약, 군집화, 분류화, 관계화, 성향, 패턴인식 등을 통하여 우리가 원하는 정보의 형태를 어떠한 방법으로 얻을 것이냐가 중요한 연구 과제이다.

1) Graduate student, Department of Statistics, Hallym University, Chuncheon, Gangwon-Do 200-702, Korea
E-mail : sklee@hallym.ac.kr
2) Professor, Department of Statistics, Hallym University, Chuncheon, Gangwon-Do 200-702, Korea. This work was supported by Hallym University, HRF-2003-32.
3) Professor, Department of Statistics, Inha Universty, 253 Younghyun-dong, Nam-gu Incheon 402-751, Korea

특히 최근에는 CRM(Customer Relationship Management)분야에서 데이터마이닝 기법을 사용하며 Witten and Frank (1999)에서 볼 수 있는 많은 활용도가 있다.

본 논문에서는 그중 하나인 고객이탈방지에 대하여 연구하여 보고자 한다. 고객이탈여부는 주로 이진변수로서 기존의 로지스틱 회귀모형이 많이 사용되었다. 반응변수 Y가 가지는 값의 집합을 $\{0,1\}$라하고 설명변수 $\boldsymbol{x}=(x_1,\cdots,x_p)^T$ 가 가지는 집합을 $\mathcal{X}$ 라 할 때, 일반적인 로지스틱 모형은 다음과 같은 형태로 표현된다.

$$P(Y=k|\boldsymbol{X}=\boldsymbol{x})=\frac{\exp\Theta(k|\boldsymbol{x})}{\exp\Theta(0|\boldsymbol{x})+\exp\Theta(1|\boldsymbol{x})},\ \boldsymbol{x}\in\mathcal{X},k\in\{0,1\} \tag{1}$$

이때 로지스틱 모형에서는 보통

$$\Theta(k|\boldsymbol{x})=\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p,k\in\{0,1\} \tag{2}$$

형태의 선형, 가법모형을 사용한다(구자용(2000)).

폴리클라스 모형은 기본적으로 함수들을 특정 기저함수로 전개하여 반응함수 $\Theta(k|\boldsymbol{x})$를 추정하게 된다. 그러므로 추정하려는 함수 $\Theta(k|\boldsymbol{x})$를 임의의 M개의 기저함수 $B_1,\cdots,B_M$로 전개하였을 때

$$\Theta(k|\boldsymbol{x})=\Theta(k|\boldsymbol{x};\beta)=\sum_{i=1}^{M}\beta_{ik}B_i(\boldsymbol{x}),\ \boldsymbol{x}\in\mathcal{X},k\in\{0,1\} \tag{3}$$

라 표현할 수 있고, 이때 기저함수 $B_1,\cdots,B_M$은

$$1,x_1,\cdots,x_p,(x_{i-t_{il}})_+,i=1,\cdots,p$$

와 이들의 텐서곱(tensor product)이 사용된다.
여기서 $(x_{i-t_{il}})_+$란 다음과 같이 표현되는 P-스플라인을 의미한다.

$$(x_i-t_{il})_+=\begin{cases}0 & \text{if} \quad x_i<t_{il}\\ (x_i-t_{il}) & \text{if} \quad x_i\geq t_{il}\end{cases} \tag{4}$$

이때 t_{il}는 x_i가 가지는 값을 크기 순서로 나열하였을 때 1번째 관측 값으로 흔히 x_i의 절단점(breal point) 또는 매듭점(knot point)라고 한다. 따라서 폴리클라스 모형에서 $P(Y=k|\boldsymbol{X}=\boldsymbol{x}\ ;\ \beta)$는 식(3)의 $\Theta(k|\boldsymbol{x}\ ;\ \beta)$를 사용한 다음과 같은 식이 된다.

$$P(Y=k|X=x\ ;\ \beta)=\frac{\exp\Theta(k|x\ ;\ \beta)}{\exp\Theta(0|x\ ;\ \beta)+\exp\Theta(1|x\ ;\ \beta)},\ x\in X,\ k\in\{0,1\} \qquad (5)$$

본 논문에서는 식(5)의 폴리클라스 모형을 확장하여 예측변수가 이산형과 연속형이 혼재 되어 있는 경우의 모형을 데이터마이닝에 적용하고자 한다.

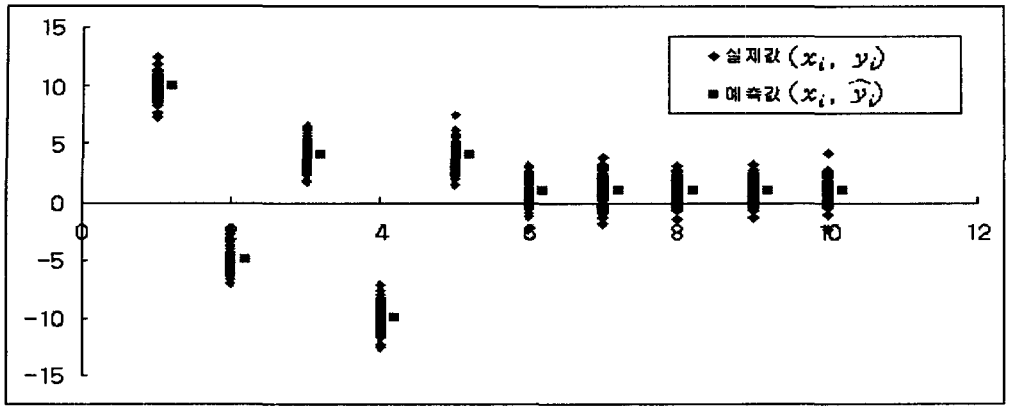
2. 이산형인 예측변수에서의 매듭점 선택

정보의 복잡화로 인하여 예측변수로 성별, 학력, 거주지와 같은 인류통계학적인 변수들이 자료에 많이 포함되어 있고 이에 따라 폴리클라스 모형에서도 예측변수가 이산형과 연속형이 혼재되어 있는 경우가 많아지게 되었다. 폴리클라스 모형의 예측변수 중 이산형인 변수 x_i의 $\{1,2,\cdots,h\}$개의 서로 다른 값을 가질 수 있다면 각각의 값을 h개의 자릿수를 갖는 이진수 형태로 변환하게 된다. 예를 들어 자료 중 고객의 현 주소를 나누는 (서울지역 = 1, 경기지역 = 2, 강원지역 = 3)으로 입력되어 있는 변수가 있다고 하면 3자리를 갖는 이진수인 $(001_{(2)}=1, 010_{(2)}=2, 011_{(2)}=3, 100_{(2)}=4, 101_{(2)}=5, 110_{(2)}=6, 111_{(2)}=7)$로 변환하여 분석하고자 한다. 이때 $(011_{(2)}=3)$이라는 값이 의미하는 바는 고객의 현 주소가 경기지역이거나 강원지역임을 의미한다. 이 때 h개의 범주를 갖는 변수라면 (2^h-1)개의 값으로 나누어 생각하게 된다.

다음은 예측변수가 이산형인 1,000개의 데이터 쌍 (x_i, y_i)를 생성하여 위에서 제시한 방법의 예측정도를 알아보자. x_i는 1에서 10사이의 이산형 일양분포에서 생성하고 y_i는 $y_i=f(x_i)+\varepsilon_i$, $\varepsilon_i \sim iid\ N(0,1)$ 로 생성하였다. 여기서 $f(x_i)$는 다음과 같다.

$$f(x_i)=\begin{cases} 10 & , x_i=1 \\ -5 & , x_i=2 \\ 4 & , x_i=3 \\ -10 & , x_i=4 \\ 4 & , x_i=5 \\ 1 & , o.w \end{cases}$$

이와 같은 모형에서 생성된 1,000개의 난수쌍 $(x_i,\ y_i)$와 이산형 예측변수에 대한 매듭점 선택 방법으로 폴리클라스 모형에 적합 시켜 본 결과인 $(x_i,\ \hat{y}_i)$를 그림 1로 표현하였다. 이 그림에서 보는 것 같이 $\hat{y}_i$는 y_i의 기대값인 $f(x_i)$를 잘 예측해 내고 있다.



<그림 1> 모의 실험을 통한 폴리클라스 모형에서의 예측값

3. 폴리클라스 데이터마이닝 모형 설계

폴리클라스 데이터마이닝 모형은 우선 입력받은 데이터를 사용자의 입력에 따라 훈련데이터와 검정데이터의 두개의 데이터로 나뉘게 된다. 여기서 훈련데이터는 실질적인 알고리즘에 의해 모형을 생성하는데 쓰이는 데이터를 말하고, 검정데이터는 훈련데이터로 만들어진 모형이 얼마나 잘 만들어진 모형인가를 검정하는데 쓰이는 데이터를 말한다. 이때 훈련데이터와 검정데이터의 비율은 약 7:3정도로 많이 사용되고 있다(장남식(1999)).

훈련데이터를 이용하여 폴리클라스 모형에서의 설명변수선택 및 제거는 다음과 같은 알고리즘을 이용하였다.

3.1 변수선택단계 : Addition stage

주어진 데이터를 이용하여 각 변수가 결과에 미치는 영향이 어느 정도인지를 판별하기 위하여 다음의 방법을 사용하였다.

1 단계 : $\widehat{\Theta}_i = \beta_0$, $\beta_0 =$ constant 의 모형을 구축한다.
2 단계 : 위 모형의 RSS(Residual Sum of Squares)값을 계산한다.
3 단계 : 기존의 모형에 각 변수와 변수의 스플라인을 추가한다.
예 : $\widehat{\Theta}_i = \beta_0 + \beta_1 (x_i - t_{il})_+$
4 단계 : 새로운 모형에서 기존모형과의 RSS의 차이 값을 계산한다.
5 단계 : RSS의 차이가 가장 큰 변수나 변수의 스플라인, 또는 각 스플라인의 곱을 매듭점으로 선정한 후 모형 및 RSS값을 갱신한다.
매듭점 선택의 순서는 다음과 같다.

a. x_i , $i=1,\cdots,p$

b. $(x_i-t_{il})_+$: 단 모형에 이미 x_i가 선택되어 있을 때 고려한다.

c. x_ix_j : 단 모형에 이미 x_i와 x_j이 선택되어 있을 때 고려한다.

d. $x_i(x_j-t_{jm})_+$: 단 모형에 이미 x_ix_j와 $(x_j-t_{jm})_+$이 선택되어 있을 때 고려한다.

e. $(x_i-t_{il})_+(x_j-t_{jm})_+$: 단 모형에 이미 $x_i(x_j-t_{jm})_+$와 $x_j(x_i-t_{il})_+$이 선택되어 있을 때 고려한다.

6 단계 : 주어진 조건이 맞을 때까지 b 단계부터 e 단계까지를 반복한다. 이 루프의탈출 조건은 다음과 같다.

탈출 조건 : GCV(generalized cross-validation)를 사용하여 GCV의 값이 감소하다가 증가하기 시작하면 루프에서 탈출한다. GCV의 계산은 다음과 같이한다.(Wahba, G., Golub, G.H. and Heath, C.G.(1979))

$$GCV=\frac{\sum e_i^2}{[1-\frac{1}{n}tr(\mathbf{H})]^2} \tag{6}$$

여기에서 $\mathbf{H}$는 hat matrix인 $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$을 의미한다.

3.2 변수제거단계 : Deletion stage

1 단계 : 변수추가단계에서 마지막으로 선정된 모형에 들어있는 매듭점 중

$$\frac{\hat{\beta}_j^2}{(\mathrm{X}'X)_{jj}^{-1}} \tag{7}$$

의 값이 최소인 매듭점을 찾는다. 이때 X는 기저들의 Design Matrix이다. 제거대상 변수의 선택 순서로 변수선택단계 중 5단계에서 선택하는 순서의 역순으로 한다.

2 단계 : 1 단계에서 찾은 매듭점을 제거하였을 때의 BIC값을 계산한다. BIC의 계산은 다음과 같이한다.

$$\mathrm{BIC}=n\log(\frac{\mathrm{RSS}}{n-M+1})+M\log n \tag{8}$$

여기서 n은 데이터의 개수, M은 기저들의 개수, RSS는 해당변인을 제거하였을때의

RSS값을 의미한다.

3 단계 : BIC의 값이 감소하면 해당 변인을 제거하는 모형으로 갱신하고 변수제거단계의 처음으로 돌아간다.
BIC의 값이 감소하지 않으면 변수제거단계를 끝낸다.

4. 적용사례

본 논문에서 사용된 자료는 최근 중요성이 높아지고 있는 고객이탈관리에 관련된 자료로서 총 2,671건으로 구성되어 있고 설명변수는 총 80개로 11개의 이산형 변수와 69개의 연속형 변수로 이루어져 있다. 서비스지속 유무를 나타내는 반응변수를 Y라 할 때

$$Y=\begin{cases}1 & (\text{고객이서비스를해지한경우}) \\ 0 & (\text{고객이서비스를지속한경우})\end{cases} \tag{9}$$

로 정의된다.

4.1 분석단계

1 단계 : 훈련데이터와 검정데이터로 분류한다. 이때 분류방법은 U(0,1)에서의 난수를 발생하여 0.3보다 낮은 값이 나오면 검정데이터로 그 이외의 값이 나오면 훈련데이터로 나누어 최종적으로 훈련데이터와 검정데이터의 비율이 약 7 :3 정도가 될 수 있도록 하였다. 본 논문에서는 약 70%에 해당하는 1,856개가 훈련데이터로, 약 30%에 해당하는 815개가 성능평가를 위한 검정데이터로 나누어졌다.

2 단계 : 1,856개의 훈련데이터로 3.1와 3.2에서 설명한 변수추가방법, 제거방법을 사용하여

$$\Theta(\hat{M}\boldsymbol{x})=\Theta(\hat{M}\boldsymbol{x};\beta)=\sum_{i=1}^{M}\beta_{ik}B_i(\boldsymbol{x}),\ \boldsymbol{x}\in X,\ k\in\{1,2\} \tag{10}$$

모형에 적합하여 $\beta_{ik}, B_i(\boldsymbol{x})$를 추정한다.

3 단계 : 815개의 검정데이터로 오판별율 및 이득률을 계산한다.

4.2 최종모형

최종모형에서 선택된 변수들은 2절의 변수 선택법에 의해서 유의한 것으로 판단된 변수들의

집합이다. 훈련데이터에 폴리클라스 알고리즘을 적용하여 추정된 계수를 사용하여 $\hat{\theta}(x)$를 얻어 본 결과

$$\begin{aligned}\hat{\theta}(x) = {} & 21.89-8.32(x_{76}-6)_+ +0.0004x_{48}+0.0008(x_{48}-15000)_+ \\ & -10.03(x_{76}-1)_+ -1.49(x_{76}-3)_+ -0.001x_{28}-0.0004(x_{48}-8064.5)_+ \\ & +0.0017(x_{27}-9900)_+ -0.0088x_{29}+0.0088(x_{29}-260)_+ +0.0283x_{42} \\ & -0.0273(x_{42}-74.4)_+ -0.000001(x_{48}-15000)_+(x_{28}-9900)_+ \\ & -0.0052(x_{28}-12900)_+\end{aligned} \tag{11}$$

가 되며 이때 해지 가능성에 대한 추정값은

$$\hat{p}(x)=\frac{\exp(\hat{\theta}(x))}{1+\exp(\hat{\theta}(x))} \tag{12}$$

로 주어진다. 이 훈련 데이터에 의한 추정값을 이용한 판별규칙 D_L은

$$(x_i-t_{il})_+=\begin{cases}1 & \text{if } \hat{p}(x) \geq 0.5 \\ 0 & \text{if } \hat{p}(x)<0.5\end{cases} \tag{13}$$

를 사용하였다.

4.3 오판별율

식(11)에서 얻어진 최종모형을 사용하여 검정데이터에 적합시켜 본 결과 오판별율은 표 1과 같았다. 이 표에서 보는 바와 같이 전체적인 오판별율은 3.681%로 총 815개의 검정데이터 중 30개의 자료에 대하여 틀린 결과를 예측했다.

<표 1> 폴리클라스 모형의 오판별율

실제값	예측값	
	0	1
0	561개(68.8344%)	10개(1.2270%)
1	20개(2.4540%)	224개(27.4847%)

4.4 이득률

각 모형에 대하여 이들의 성능(performance)을 비교하기 위하여 이득률(gain)을 도입하고자 한다. 특정 판별규칙 D에 의하여 구한 사후확률 모형, 즉 $P(Y=1|\boldsymbol{X}=\boldsymbol{x})$에 대한 추정규칙을 $\hat{S}^D(\boldsymbol{x})$로 나타내고 기존 고객의 스코어를

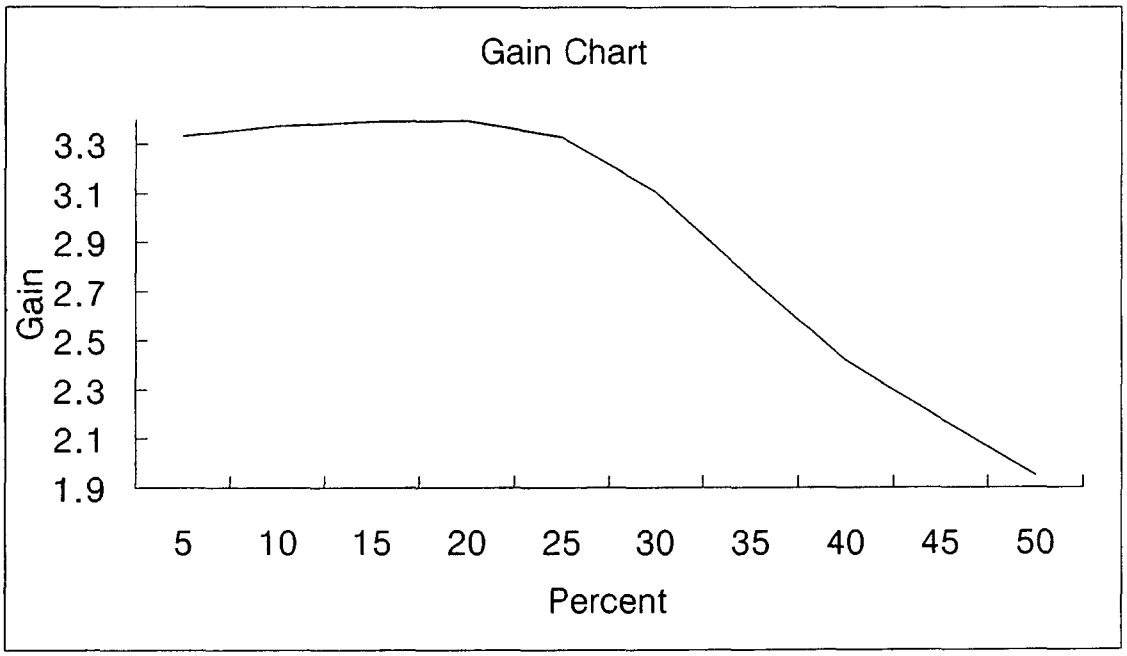
$$s_i^D = \hat{S}^D(\boldsymbol{x}_i),\ 1 \le i \le n \tag{14}$$

이라 하면, 이때 전체 데이터에서 실제로 관심의 대상이 되는 데이터 수를 N_1라하고 s_i^D의 크기 순서로 나열하였을 경우 상위 $p\times 100\%$ 데이터 중에서 실제로 관심이 되는 데이터의 수를 N_p^D라 하면 이득률은

$$\text{이득률}(p, D) = \frac{N_p^D}{N_1 \times p},\ 0 < p < 1 \tag{15}$$

로 정의한다. 판별규칙 D가 아무런 판별력을 갖지 못하는 경우 이득률 p에 상관없이 1에 가까워지며 반대로 판별력이 좋을 경우 그 값이 커지게 된다(구자용(2000)).

$p = 0.05, 0.10, \cdots, 0.50$에 대하여 D_L에 대한 이득률을 구하여 그래프로 표현하면 그림 2와 같다.



<그림 2> 폴리클라스 방법으로 생성한 모형의 상위 50%에 대한 이득률

$p \leq 0.4$인 경우 D_L을 이용하여 구매 가능성이 높은 순서로 하면 상위 $(100 \times p)\%$는 실제로 이득률이 높게 나왔지만 $p > 0.4$인 경우는 오판별 되는 경우가 생겨 이득률이 감소하기 시작한다. 그러나 구매자 예측관점에서 보면 p가 적은 경우(보통 $p < 0.2$)에 이득률이 높은 의미를 가지므로 이득률 관점에서 D_L의 성능이 좋은 것으로 판단된다.

4.5 로지스틱 모형과 폴리클라스 모형의 비교

폴리클라스 모형의 성능을 알아보기 위하여 기존의 로지스틱 모형과 오판별율을 비교하여 보았다. 이 비교에 사용된 자료는 독일신용평가 자료와 4절에서 사용하였던 자료이며, 동일한 훈련 데이터로 폴리클라스 모형과 로지스틱 모형을 만들고 검정데이터를 이용하여 두 모형의 성능을 비교하여 보는 실험을 100회 반복하였다. 총 100회의 실험에서 폴리클라스 모형과 로지스틱 모형의 오판별율을 각각 계산하여 두 가지 모형 중 어느 모형의 오판별율이 더 작은지 그 개수를 세어 보았다. 이때 폴리클라스 모형의 변수선택에서 스플라인 기저가 포함되었는지를 확인하여 스플라인 기저가 포함된 경우와 포함되지 않았을 경우로 구분하여 살펴보았는데 폴리클라스 모형에서 스플라인 기저가 포함되지 않았을 때는 기존의 로지스틱 모형과 같은 오판별율을 가졌고 스플라인 기저가 포함되었을 경우 로지스틱 모형보다 폴리클라스 모형이 오판별율이 작은 것으로 관찰되었다. 독일신용평가 자료의 결과(표 2)보다 모 이동통신회사의 자료의 결과(표 3)가 스플라인이 기저로서 포함된 경우가 많았고 결과적으로 오판별율은 100회의 실험 중 98회가 폴리클라스 모형의 오판별율이 더 낮게 나와 폴리클라스 모형이 로지스틱 모형보다 오판별율에서 우수함을 보여주었다.

<표 2> 독일신용평가자료에서의 로지스틱 모형과 폴리클라스 모형의 비교

Logistic	Polyclass	스플라인 기저의 유무
37		없음
4	59	있음

<표 3> 모 이동통신자료에서의 로지스틱 모형과 폴리클라스 모형의 비교

Logistic	Polyclass	스플라인 기저의 유무
0		없음
2	98	있음

5. 결론

일반적인 데이터마이닝 작업은 반응변수가 연속형이면 회귀분석 모형을, 범주형이면 판별분석 모형으로 나누어 수행하는데 반면, 폴리클라스 모형은 이 두 가지 모든 경우와 연속형 변수와 범주형 변수가 혼재되어 있는 경우에도 적용 가능하다. 따라서 폴리클라스 모형을 이용한 데이터마

이닝은 다양한 영역에서 사용될 수 있다는 장점을 지니고 있다. 또한 폴리클라스 모형은 Tree 모형의 장점인 해석력과 Neural Network 모형의 장점인 예측력을 동시에 수용하며, Neural Network 모형과 비슷하거나 우월한 예측 결과를 보여주는 경우도 있다.

참고문헌

[1] 구자용 (2000) "기저함수 방법론", 2000년도 한국통계학회 추계학술논문발표회 초청강좌, pp. 47-49.

[2] 구자용, 박헌진, 최대우 (2000) "데이터마이닝에서의 폴리클라스", 응용통계연구 제 13권 2호, pp. 489-503.

[3] 장남식 (1999) "데이터마이닝", 대청.

[4] Chambers, J.H. and Hastie, T.J. (1991) *Statistical Model in S*, 2/e, Wadsworth & Brooks, New York.

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

[6] Witten,Ian H. and Frank, E. (1999) *Data Mining*, Morgon Kaufmann publishers, San Francisco.

[7] McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, 2/e, Chapman and Hall, London.

[8] Press, William H. (1992) *Numerical Recipes in C*, 2/e, Cambridge University Press, W.H.

[9] Wahba, G., Golub, G.H. and Heath, C.G.(1979) *Generalized Cross Validation as Method for Choosing a Good Ridge Parameter*, Technometrics, **21**, 215-223.

[2004년 1월 접수, 2004년 4월 채택]