The Korean Communications
in Statistics Vol. 12 No. 3, 2005
pp. 659-672

Discretization Method Based on Quantiles for Variable Selection Using Mutual Information[1)]

Woon Ock Cha[2)] and Moon Yul Huh[3)]

Abstract

This paper evaluates discretization of continuous variables to select relevant variables for supervised learning using mutual information. Three discretization methods, MDL, Histogram and 4-Intervals are considered. The process of discretization and variable subset selection is evaluated according to the classification accuracies with the 6 real data sets of UCI databases. Results show that 4-Interval discretization method based on quantiles, is robust and efficient for variable selection process. We also visually evaluate the appropriateness of the selected subset of variables.

Keywords : variable selection, discretization, mutual information, data visualization

1. 서론

대용량 데이터베이스에 대한 지도학습에서, 분류를 위해 모든 예측변수를 다 사용하는 경우에는 부적절하고 불필요한 변수 때문에 분류 정확도가 떨어질 수 있고 또한 계산량은 엄청나게 된다. 따라서 목표변수에 영향을 미치는 중요한 변수를 선택하는 것은 분류에서 매우 중요한 문제이며, 통계학 분야에서도 많은 연구가 이루어졌고 패턴인식 분야에서도 많이 다루어진 연구주제이다 (Devijver and Kittler, 1982, Miller, 1990, Cha and Huh, 2003). 중요한 변수를 선택하는 문제는 목표변수와 관련이 많은 예측변수에 대해 중요한 변수 순으로 순위를 정하거나, 목표변수에 영향을 미치는 가장 적절한 변수들의 부분집합을 찾는 탐색문제이며, 탐색방법과 변수들에 대한 평가방법에 따라 다양한 변수선택방법들이 연구되었다. 변수선택방법은 필터 (filter) 와 포장 (wrapper) 방법으로 분류할 수 있는데, 필터방법은 변수들을 평가할 때 학습 알고리즘과는 독립적으로 데이터가 가지고 있는 성질을 이용하는 방법이고, 포장방법은 학습 알고리즘에 의한 분류 정

1) This research was financially supported by Hansung University in the year of 2005.
2) Professor, Division of Computer Engineering, Hansung University, Seoul, 136-792, Korea
E-mail : wcha@hansung.ac.kr
3) Professor, Department of Statistics, Sungkyunkwan University, Seoul, 110-745, Korea

확도를 사용하여 변수들을 평가하는 방법이다. 필터방법에서 평가에 사용되는 척도로는 종속성, 거리, 정보, 일치성 등이 있다 (Das and Liu, 1997, Liu and Motoda, 1998). 본 논문에서는 변수선택을 위해 필터방법과 포장방법을 같이 고려한다.

상호정보 (mutual information : MI)는 두 집단 간의 상호의존성 (interdependency) 을 평가할 수 있는 척도로서, 두 집단이 서로 독립적이면 MI는 0이고, 두 집단이 서로 종속성을 가지고 있으면 이 값은 커지게 된다. 특히 MI는 관련되어 있는 변수들의 형식이 연속형이거나 이산형에 관계가 없을 뿐만 아니라 비선형적인 관계에서도 사용할 수 있기 때문에 예측변수의 목표변수에 대한 예측정도를 나타내주는 척도로 최근 많이 연구되고 있다. 본 논문에서는 필터방법의 평가척도로서 정보를 이용하는 상호정보방법을 사용하고, 중요변수의 부분집합을 결정할 때는 포장방법을 사용한다. 예측변수가 연속형 값을 갖는 경우에 데이터로부터 MI를 직접 구하기 위해서는 Parzen 윈도우 (Parzen, 1962) 를 사용하는 방법 등이 제안되었는데, 이 방법을 사용하기 위해서는 결합밀도함수를 추정해야 하는 어려운 문제점이 있다. 이러한 이유 때문에 연속형 데이터를 먼저 이산화시키고 이산형 데이터에 대한 MI를 사용하는 것이 일반적인 접근방법이다.

본 논문의 목적은, 예측변수가 연속형이고 목표변수가 범주형인 데이터이스로부터 예측변수 값을 먼저 이산화시킨 후 MI에 의해 중요한 변수를 선택하는 문제에서, 여러 가지 이산화방법에 따른 분류결과를 비교하고 좋은 결과를 얻을 수 있는 효율적인 이산화방법을 찾고자 하는 것이다. 본 논문의 구성은 다음과 같다. 2장에서는 MI를 이용하는 변수선택 알고리즘을 제안하고, 3장에서는 세 가지 이산화방법에 대해 정리하였다. 4장에서는 본 연구에서의 실험방법과 실험에 사용하는 데이터베이스, 실험결과를 기술하였고 5장에 분석 및 평가를 기술하였다. 실험결과에 의하면 복잡한 이산화방법을 사용하는 것보다 간단한 4-구간 방법을 사용하여 이산화를 하고 중요한 변수를 선택하여 분류를 수행하면 좋은 분류결과를 얻을 수 있게 된다.

2. MI에 의한 중요변수 부분집합 결정방법

데이터베이스로부터 중요한 변수들의 부분집합을 구하는 과정에서 가장 확실한 방법은 p개 변수에 대한 2^p개의 부분집합을 고려하는 것이다. 그러나 p가 10개 정도만 되어도 이를 위한 과정은 불가능해진다. 따라서 적정한 기준에 의해 중요변수의 부분집합을 구하는 것이 필요하게 된다. 중요한 변수들로만 이루어지도록 데이터베이스를 축소시키면 분류과정이 보다 빨리 효율적으로 수행될 수 있으며, 이로부터 생성된 분류를 위한 지식의 형태는 보다 간결하고 이해하기 쉽게 된다.

2.1 MI

MI는 변수들이 만족시켜야 하는 성질에 대한 특별한 가정을 하지 않고, 변수들 간의 선형 종속성뿐만이 아니라 일반적인 종속성을 측정할 수 있는 방법이다 (Tourassi, Frederick, Markey and Floyd, Jr., 2001).

X를 예측변수, Y를 목표변수라 하면, X와 Y사이의 MI I(X; Y)는 다음과 같이 정의한다.

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

$$= H(Y) - H(Y|X)$$

여기에서 $H(X)$는 변수 X의 불확실성 (uncertainty) 을 나타내는 엔트로피 (entropy) 이며, X가 이산형 변수일 경우에는

$$H(X) = -\sum_{x} p(x) \log p(x)$$

로 정의한다. 또한 $H(Y|X)$ 는 다음과 같이 정의한다.

$$H(Y|X) = -\sum_{y} p(y|x) \log p(y|x)$$

여기에서 $p(x) = P[X = x]$ 는 변수 X의 확률분포함수이고, $p(y|x) = P[Y = y|X = x]$는 X가 주어졌을 때 Y의 조건부확률이다. MI는 변수 X가 알려졌을 때 Y의 불확실성을 얼마나 감소시킬 수 있는지를 측정하는 것이며, 예측변수가 연속형일 경우에는 계산이 복잡하여 근사식을 사용하거나 미리 이산화를 시킨 후 다음 식에 의해 MI $I(X; Y)$를 계산할 수 있다.

$$I(X; Y) = \sum_{x}\sum_{y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

이 식에서 $p(y) = P[Y = y]$는 변수 Y의 확률분포함수이고 $p(x,y) = P[X = x, Y = y]$는 결합확률분포함수이다.

변수 벡터 $(X_1, X_2, \cdots, X_p)$가 제공하는 정보가 목표변수 Y의 불확실성을 얼마나 줄여줄 수 있는지를 측정하는, 예측변수 $(X_1, X_2, \cdots, X_p)$와 목표변수 Y의 결합 MI (joint mutual information : JMI)는 다음 식과 같다(Cover and Thomas, 1991).

$$I(X_1, X_2, \cdots, X_p; Y) = \sum_{i=1}^{p} I(X_i; Y|X_{i-1}, X_{i-2}, \cdots, X_1)$$

여기에서,

$$I(X_i; Y|X_{i-1}, X_{i-2}, \cdots, X_1) = H(X_i|X_{i-1}, X_{i-2}, \cdots, X_1) + H(Y) - H(X_i, Y|X_{i-1}, X_{i-2}, \cdots, X_1)$$

이다.

연속형 변수가 가지는 값을 이산화 시킬 때 구간의 수가 많아지면 MI가 작아지는 경향이 있다. 에에 대한 근거는 4절의 실험 및 실험결과에서 다시 논의하기로 한다.

2.2 중요변수 부분집합 결정방법

MI를 이용하여 p개의 예측변수 $X_i, i=1, \cdots, p$ 중에서 목표변수 Y에 영향을 미치는 중요한 변수를 찾기 위해서는, p개의 변수로부터 생각할 수 있는 모든 가능한 2^p개의 부분집합에 대해 JMI를 계산해서 가장 큰 값을 가지는 변수들의 부분집합을 찾아야 한다. 그러나 변수 수가 많은 경우에는 계산상의 문제를 초래하게 된다. 본 논문에서는 크기가 q이하인 중요변수 부분집합 S를 결정하기 위해서 Battiti(1994)의 방법에 바탕을 둔 발견론적 알고리즘을 제안한다.

$I \leftarrow \{1, 2, \cdots, p\}$;
$S \leftarrow \varnothing$;
$I(\varnothing; X_i) \leftarrow 0$;
for m = 1 **to** q
{
$j = \text{indexMax}_{i \in I}\{I(X_i; Y) - \beta \sum_{k \in S} I(X_k; X_i)\}$
$S \leftarrow S \cup \{j\}$; $I \leftarrow I - \{j\}$;
if (Acc(S) $\geq$ Acc(전체 데이터베이스)) **break**;
if (Acc(S) $\leq$ Acc(S-{j})) {$S \leftarrow S$-{j}; **break**;}
}

여기에서 β는 다음에 선택 될 후보변수와 이미 선택된 변수 사이의 상호의존성을 제어하는 매개변수이다. 만일 β가 0이면 목표변수와 후보변수 사이의 상호정보 값만 고려해서 변수를 선택하는 것이고, β가 0보다 크면 이 값에서 후보변수와 이미 선택된 변수 사이의 상호정보의 합을 β배 만큼 제외한 값을 이용해서 다음 변수를 선택하는 것이다. 일반적인 분류문제에서는 β가 0.5 ~ 1 사이의 값을 가지는 것이 적합하다(Battiti, 1994). 이 알고리즘을 적용하면 이미 결정된 중요변수 부분집합에 속하는 변수와 상호의존성이 큰 변수는 부분집합의 다음 원소로 결정될 가능성이 적어지게 된다.

또, Acc(S)는 S에 속하는 변수번호 (index)를 가지는 변수들과 목표변수로 이루어진 데이터 파일에 대한 분류정확도, Acc(전체 데이터베이스)는 전체 데이터베이스에 대한 분류정확도로서, 데이터 파일에 로지스틱(logistic)이나 나이브 베이즈(Naive Bayes)와 같은 분류모형을 적용하여 10-층 교차타당성(10-fold cross validation)방법으로 구하는 것이다. 미리 정의한 크기 q가 되기 전에 부분집합의 정확도가 전체 데이터베이스를 다 사용한 경우의 정확도보다 좋거나, 변수 하나를 더 추가한 경우의 정확도가 추가하기 전의 정확도 보다 향상되지 않는 경우는 크기 q미만의 부분집합을 구하고, 그렇지 않으면 크기 q인 S를 구한다. 이렇게 해서 만들어진 집합 S의 원소들은 중요변수 부분집합에 해당하는 변수번호 집합이다. 그리고 S에 속해있는 원소의 순위는 선택된 부분집합들의 순서를 나타낸다. 예를 들어 S = {3, 4, 1, 2} 이면, 처음에 선택된 변수 부분집합은 변수번호 {3}으로 구성된 것이고, 두 번째 선택된 변수 부분집합은 {3, 4}, 세 번째 선택된 변수 부분

집합은 {3, 4, 1}, 네 번째 선택된 변수 부분집합은 {3, 4, 1, 2}를 나타낸다.

3. 이산화방법

연속형 변수 값에 대한 이산화 방법은 전역적 이산화와 지역적 이산화로 나눌 수 있는데, 전역적 이산화는 분류모형을 사용하기 이전에 전체 데이터에 대해 수행하며, 지역적 이산화는 결정나무의 생성과정에서와 같이 각 노드에서 그 노드에 도달한 데이터만을 이용하여 부분적으로 이산화를 수행한다. 전역적 이산화 방법은 분류의 전처리 과정으로 유용하게 사용할 수 있으며 본 논문에서는 전역적 이산화 방법인 MDL (Minimum Description Length), 히스토그램, 4-구간 방법만 고려한다.

(1) MDL 방법

전역적 이산화를 위해 많이 사용하는 것은 Gini 계수 (Breiman et al, 1984), 정보획득 (information gain) 등 불순도 척도 (impurity measure) 에 바탕을 둔 방법이다. MDL은 정보획득에 바탕을 둔 것으로 Fayyad와 Irani(1992)가 제안한 방법이다.

이 방법에서는 분할점 (split point)을 선택하기 위해 후보 분할점들의 평균 클래스 엔트로피 (average class entropy) $E(X, T; D)$를 이용한다.

D를 데이터집합, X를 연속형 변수, T를 후보 분할점, 그리고 D_1, D_2는 D를 T로 분할한 경우의 부분구간의 데이터집합이라 할 때, D의 클래스 엔트로피 함수 $H(D)$와 평균 클래스 엔트로피 $E(X, T; D)$는 다음과 같다.

$$H(D) = -\sum_{i=1}^{k} p(C_i, D) \log\,(p(C_i, D)),$$

여기에서 목표변수 Y의 i번 째 클래스를 C_i라 할 때, $p(C_i, D)$ 는 D가 클래스 C_i에 속할 확률이고 k는 클래스의 개수이다.

$$E(X, T; D) = \frac{|D_1|}{n} H(D_1) + \frac{|D_2|}{n} H(D_2),$$

여기에서 $n = |D|$이다.

이 방법에서는 다중 분할점을 생성하기 위해, 연속형 변수 X에 대한 모든 가능한 분할점 중 평균 클래스 엔트로피를 가장 최소화하는 T_C 를 분할점으로 선택하여 구간을 두 개로 분할하고 이 방법을 각각의 분할 구간에 재귀적으로 적용하여 특정 조건이 만족 될 때까지 이진분할을 계속한다. 이와 같은 이진분할은 다음과 같은 조건을 가질 때 중지한다(Fayyad와 Irani(1992)).

$$\text{정보획득} > \frac{\log(n-1)}{n} + \frac{\log(3^k - 2) - kE + k_1 E_1 + k_2 E_2}{n}$$

여기서 log는 밑을 2로 하며, $E = H(D)$, $E_1 = H(D_1)$, $E_2 = H(D_2)$ 를 나타내고 k_1, k_2

는 각 분할에 속해 있는 클래스의 개수이다. 이 방법에서는 연속형 변수값을 이산화 시킬 때 예측 변수 자체만 고려하는 것이 아니라 목표변수의 정보까지 같이 고려하는 것이다. MDL방법으로 이산화를 수행하면 변수마다 구간 수가 달라질 수 있다.

(2) 히스토그램 방법 (HIS)

이 방법에서는 변수가 정규분포를 따르는 경우와 그렇지 않은 경우의 구간의 개수를 수정된 Sturge의 규칙에 따라 다음과 같이 정한다(Vernables and Ripley, 1994).

$$\text{구간의 개수} = \log_2 n + 1 , \qquad \text{정규분포를 따르는 경우}$$

$$\text{구간의 개수} = \log_2 n + 1 + \log_2(1 + |x|\sqrt{n/6}) , \qquad \text{정규분포를 따르지 않는 경우}$$

여기에서 n은 관측값의 크기이고 x 는 첨도의 추정값이다. 구간을 정하기 위해 적절한 방법에 의해 평균 μ와 표준편차 σ를 추정하고, [μ-2σ, μ+2σ] 구간을 정해진 구간의 개수만큼의 동일한 구간으로 나눈다. 만약 [μ-2σ, μ+2σ] 바깥에 관측값이 있는 경우에는 구간보다 작은 값에 대해 하나의 구간, 큰 값에 대해 또 하나의 구간을 만든다.

이 방법을 적용하면 각 변수마다 구간 수의 차이가 있을 수 있다. 또한 관측값의 수가 많아질수록 구간의 수가 늘어나기 때문에 일관된 MI 추정량이 제공되지 못하고, 구간 수가 많아지면 분할표의 어떤 구간에는 관측값이 적거나 없을 수가 있어 MI를 계산하는데 문제가 된다.

(3) 4-구간 방법

각 구간이 동일한 수의 관측값을 가지도록 구간을 나누는 방법은 Bonnlander 와 Weigend (1994) 가 연구한 바 있다. Bonnlander 와 Weigend 가 제시한 방법은 구간수를 결정할 때 목표변수의 정보를 사용하며 구간수의 추정 과정이 복잡하다. 본 논문에서는 목표변수의 정보를 이용하지 않는 상자도형 (Boxplot) 에 기본을 둔 방법을 제안한다. 상자도형의 경우, 최대 4 개의 구간을 갖고, 양 측에 모두 특이값이 존재할 때는 최대 6개의 구간 수를 갖는다. 그러나 특이값에 해당하는 구간에는 관측값이 몇 개 속하지 않기 때문에 이 구간에서의 확률 추정이 불안해 질 수 있다. 또한 구간 수는 모든 변수에서 동일하게 정해져야 변수들의 상호 비교가 의미를 갖기 때문에, 본 논문에서는 변수가 갖는 값에 대한 하사분위수, 중앙값, 상사분위수에 따라 구간을 4개로 나누고, 각 상자도형에서 양측의 특이값에 해당하는 구간을 상-하 사분위수 구간에 포함시킨다. 이 방법을 적용하면 모든 변수에 대한 이산화가 4개의 구간으로 통일되므로 일관성이 유지되고 MI 추정이 표준화되기 때문에 상대평가가 용이해지는 장점이 있다. 또한 구간의 수가 4개로 제한되어있기 때문에 관측값의 크기가 크지 않더라도 각 구간에 속하는 관측값의 수가 어느 정도의 수준을 유지하므로 해당 구간에서의 확률 추정에 대한 신뢰도가 높아진다.

4. 실험 및 실험결과

4.1 실험설계

본 연구에서는 다음과 같은 방법으로 실험하였다.

1. 연속형 데이터베이스를 MDL, 히스토그램, 4-구간 방법으로 이산화시킨다.

2. 각각의 방법으로 이산화시킨 데이터베이스에 대해 목표변수에 대한 MI가 큰 순서로 예측변수의 순위를 구한다.
3. 2.2의 알고리즘을 사용하여 중요변수의 부분집합 S를 구한다. β값을 0.5부터 1까지 변화시켜 가면서 실험해 본 결과 큰 차이가 없었으므로 β = 0.7을 사용하고 분류 정확도는 로지스틱 모형(LOGISTIC)과 나이브 베이즈(NB) 모형으로 구한다. 또 여러 가지 실제 데이터베이스를 사용하여 실험해 본 결과 예측변수를 4개 정도 사용하였을 때 분류결과가 충분히 좋게 나타났기 때문에 S의 최대크기를 4로 한다.

각 단계에서의 실험을 위해, 히스토그램 방법으로 이산형 데이터베이스를 생성하는 것과 MI의 계산 및 2.2 방법에서 중요변수 부분집합의 후보변수를 구하는 것은 R 프로그램을 사용하였다. (Ihaka and gentleman, 1967). MDL 방법과 4-구간 방법에 의한 이산형 데이터베이스는 DAVIS를 사용하여 생성하였고 (Huh, 2005), 분류 정확도는 WEKA의 *Experiment*의 *logistic, NaiveBayes* 분류모형을 적용하여 구하였다 (Witten and Frank, 1999).

4.2 사용 데이터베이스

실험을 위하여, 모든 예측변수들이 연속형 값을 가지고 있고 목표변수는 범주형이며 결측치를 포함하지 않는 데이터베이스를 UCI 창고 (Merz and Murphy, 1996) 에서 구하였다. 본 실험에서는 붓꽃 데이터베이스 (IRIS), 갑상선질환 데이터베이스 (THYROID), 피마 인디언 당뇨병 데이터베이스(Pima Indians Diabetes : PIMA), 포도주인식 데이터베이스 (Wine Recognition : WINE), 이미지 분리 데이터베이스 (Image Segmentation : IMAGE), 위스콘신 진단 데이터베이스 (Wisconsin Diagnostic : WDBC)의 6개를 사용하였다. 예측변수의 수와 클래스의 개수 및 데이터베이스의 크기는 다음 표와 같다.

<표 1> 사용 데이터베이스

데이터베이스	예측변수 수	목표변수의 클래스 수	크기
IRIS	4	3	150
THYROID	5	3	215
PIMA	8	2	768
WINE	13	3	178
IMAGE	18	7	210
WDBC	30	2	569

4.3 실험결과

4.3.1 이산화방법에 따른 변수별 MI 순위와 중요변수 부분집합 S

세 가지 방법으로 이산화 시킨 데이터베이스에 대해 변수별 MI 순위와 중요변수 부분집합 S를 구해 다음 <표 2> ~ <표 7>에 정리하였다.

<표 2> IRIS에 대한 MI 순위와 S 집합

	MI 순위	LOGISTIC	NB
MDL	4, 3, 1, 2	{4}	{4, 3}
HIS	4, 3, 1, 2	{4}	{4, 3}
4-구간	3, 4, 1, 2	{3, 4}	{3}

<표 3> THYROID에 대한 MI 순위와 S 집합

	MI 순위	LOGISTIC	NB
MDL	2, 5, 1, 3, 4	{2, 5}	{2, 5}
HIS	2, 4, 1, 5, 3	{2, 4, 5, 3}	{2, 4, 5, 3}
4-구간	2, 5, 3, 1, 4	{2, 5, 1}	{2, 5, 1}

<표 4 > PIMA 에 대한 MI 순위와 S 집합

	MI 순위	LOGISTIC	NB
MDL	2, 6, 8, 5, 1, 7, 3, 4	{2, 6, 8, 7}	{2, 6}
HIS	2, 6, 5. 8, 3, 4, 7, 1	{2, 1, 7}	{2, 1, 7}
4-구간	2, 6, 8, 5, 1, 4, 7, 3	{2, 6, 8, 7}	{2, 6}

<표 5> WINE에 대한 MI 순위와 S 집합

	MI 순위	LOGISTIC	NB
MDL	7, 13, 10, 12, 11, 1, 6, 2, 4, 9	{7, 13, 10, 4}	{7, 13, 10, 4}
HIS	13, 7, 12, 10, 11, 1, 4, 2, 6, 5	{13, 7, 2, 12}	{13, 7, 2, 12}
4-구간	7, 13, 10, 12, 1, 11, 6, 9, 2, 4	{7, 1, 11, 13}	{7, 1, 11, 13}

<표 6> IMAGE에 대한 MI 순위와 S 집합

	MI 순위	LOGISTIC	NB
MDL	18, 9, 16, 12, 10, 11, 15, 17, 13, 14	{18, 16, 17, 15}	{18, 16, 17}
HIS	2, 15, 16, 10, 18, 13, 14, 11, 12, 9	{2, 18, 10}	{2, 18, 10}
4-구간	18, 16, 12, 9, 10, 11, 17, 14, 2, 15	{18, 12, 2}	{18, 12, 2}

<표 7> WDBC에 대한 MI 순위와 S 집합

	MI 순위	LOGISTIC	NB
MDL	23, 24, 21, 28, 8, 3, 4, 1, 7, 14	{23, 28}	{23, 28}
HIS	24, 23, 4, 13, 8, 14, 3, 11, 21, 28	{24, 11, 8}	{24, 11, 8, 28}
4-구간	21, 23, 24, 28, 8, 3, 4, 1, 7, 14	{21, 28}	{21, 28}

붓꽃 자료의 경우 변수별 MI 순위와 S집합 모두 MDL 방법과 히스토그램 방법에서는 동일하였으나, 4-구간 방법에서는 조금 다르게 나타났다. 그 외 여러 데이터베이스에서 MI 순위와 S집합이 첫 번째, 두 번째 변수정도를 제외하고는 서로 다르게 나타남을 알 수 있고, 이는 변수들 간의 상호의존성이 많이 존재한다는 것을 의미한다.

또 이산화를 시킬 때 구간의 수가 많아지면 MI의 크기가 작아지는 경향을 실험결과로 확인할 수 있었다. PIMA, WDBC 데이터베이스의 경우를 예로 들면, 각 이산화방법에 따른 특정 변수의 MI 값은 다음과 같다. 우선 PIMA 데이터베이스의 경우는,

MDL(2 ; 4) = 0.19008255
HIS(2 ; 28) = 0.05569810
4-구간(2 ; 4) = 0.17037840

이다. 여기에서 MDL(2 ; 4)는 변수 2를 MDL 방법으로 이산화하여 4개의 구간으로 나누었을 때의 MI를 의미하고, HIS(2 ; 28), 4-구간(2 ; 4)도 각 방법으로 변수 2를 28개, 4개의 구간으로 나누었을 때의 MI를 나타낸다. WDBC 데이터베이스의 경우는,

MDL(23 ; 4) = 0.68504369
HIS(23 ; 27) = 0.118144920
4-구간(23 ; 4) = 0.618434932

이다.

4.3.2 이산화방법에 따른 변수선택과 분류정확도

세 가지 이산화 방법에 대해, S집합의 크기와 선택된 변수들로만 이루어진 축소된 데이터베이스에 대한 로지스틱, 나이브 베이즈 분류모형의 정확도를 다음 <표 8> ~ <표 13>에 정리하였다. 표의 마지막 행은 전체 데이터베이스의 변수 수와 분류 정확도를 나타낸다. 선택된 변수 수 옆의 %는 전체 변수 수에 대한 백분율을 나타내고, 분류정확도 옆의 괄호 안은 전체 데이터베이스에 대한 정확도를 100으로 나타낼 때 해당 정확도를 의미한다.

<표 8> IRIS에 대한 변수선택과 분류정확도

이산화방법	LOGISTIC		NB	
	S 집합의 크기	분류정확도	S 집합의 크기	분류정확도
MDL	1(25%)	96.00(100)	2(50%)	96.00
HIS	1	96.00	2	96.00
4-구간	2	96.00	1	96.67(100.7)
FULL data	4(100%)	96.00(100)	4(100%)	96.00(100)

<표 9> THYROID에 대한 변수선택과 분류정확도

이산화방법	LOGISTIC		NB	
	S 집합의 크기	분류정확도	S 집합의 크기	분류정확도
MDL	2(40%)	96.28(99.5)	2	97.66(100.96)
HIS	4(80%)	96.73(99.9)	4	96.73(99.9)
4-구간	3(60%)	97.21(100.5)	3	97.71(101.0)
FULL data	5(100%)	96.77(100)	5(100%)	96.73(100)

<표 10> PIMA에 대한 변수선택과 분류정확도

이산화방법	LOGISTIC		NB	
	S 집합의 크기	분류정확도	S 집합의 크기	분류정확도
MDL	4(50%)	77.74(100.7)	2(25%)	76.43(100.2)
HIS	3(37.5%)	75.91(98.3)	3	75.92(99.5)
4-구간	4	77.74(100.7)	2	76.43(100.2)
FULL data	8(100%)	77.22(100)	8(100%)	76.31(100)

<표 11> WINE에 대한 변수선택과 분류정확도

이산화방법	LOGISTIC		NB	
	S 집합의 크기	분류정확도	S 집합의 크기	분류정확도
MDL	4(30.8%)	93.79(96.5)	4	94.90(98.2)
HIS	4	93.86(96.6)	4	91.12(94.3)
4-구간	4	96.05(98.8)	4	97.22(100.6)
FULL data	13(100%)	97.19(100)	13(100%)	96.63(100)

<표 12> IMAGE에 대한 변수선택과 분류정확도

이산화방법	LOGISTIC		NB	
	S 집합의 크기	분류정확도	S 집합의 크기	분류정확도
MDL	4(22.2%)	87.14(105.8)	3	78.10(100.6)
HIS	3(16.7%)	85.24(103.5)	3	83.81(108.0)
4-구간	3	85.71(104.0)	3	83.81(108.0)
FULL data	18(100%)	82.38(100)	18(100%)	77.62(100)

<표 13> WDBC에 대한 변수선택과 분류정확도

이산화방법	LOGISTIC		NB	
	S 집합의 크기	분류정확도	S 집합의 크기	분류정확도
MDL	2(6.7%)	94.03(100.6)	2	94.21(101.3)
HIS	3(10.0%)	93.50(100.0)	4(13.3%)	94.03(101.1)
4-구간	2	94.38(100.9)	2	94.73(101.9)
FULL data	30(100%)	93.50(100)	30(100%)	92.98(100)

예측변수 4개미만을 사용하여 분류했을 경우 전체 변수를 다 사용한 분류 결과보다 로지스틱, 나이브 베이즈 모형에서 거의 더 좋게 나타났다. 또 IRIS와 같이 분별력이 좋은 데이터베이스의 경우에는 이산화방법이 변수선택에 별 영향을 미치지 않는다는 것과, 4-구간 방법으로 이산화시킨 후 변수를 선택하는 경우 이산화방법은 매우 간단하지만 다른 방법에 비해 더 좋은 결과를 얻을 수 있음을 알 수 있다.

4.3.3 데이터 탐색

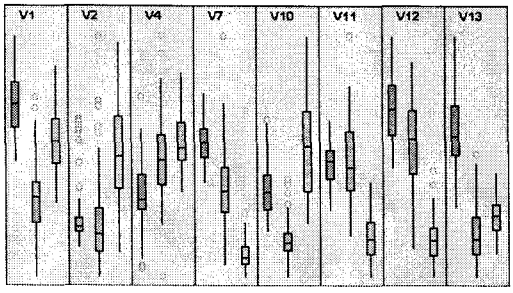
DAVIS를 이용하여 WINE, IMAGE, WDBC 데이터베이스에 대해 그린 상자도형과 상관도(Scatter Diagram)를 <그림 1> ~ <그림 6>에 나타내고 4.3.1의 결과와 비교 분석하였다.

(1) WINE 데이터베이스

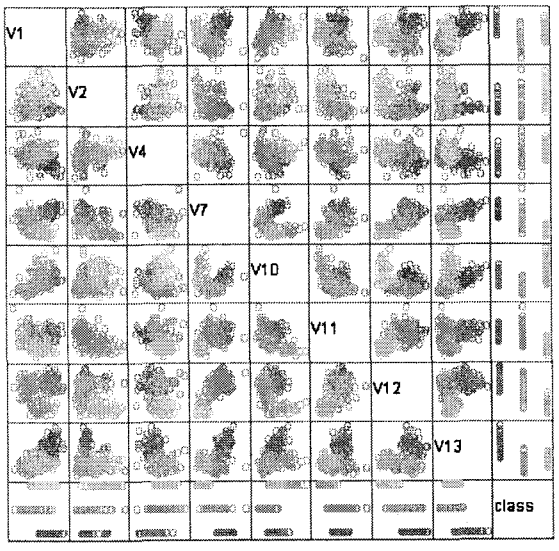
세 가지 이산화 방법에 따라 로지스틱 모형으로 구한 집합 S에 포함된 v1, v2, v4, v7, v10, v11, v12, v13에 대한 상자도형을 보면, v7, v1이 가장 중요한 변수이고 그 다음 v13이 중요한 변수임을 알 수 있다. 상관도를 보면 v7과 v1, v7과 v11은 서로 연관관계가 약하다. 따라서 4-구간 방법으로 이산화시켜 얻어진 {v7, v1, v11, v13}이 다른 이산화방법의 결과보다 중요한 변수를 더 잘 선택한 것임을 알 수 있다.

(2) IMAGE 데이터베이스

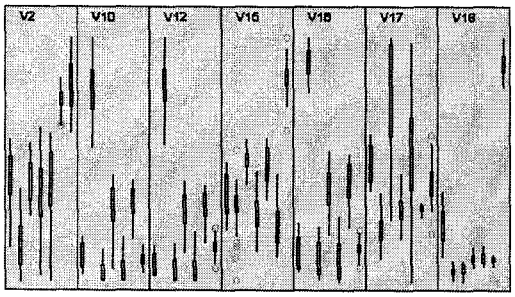
세 가지 이산화 방법에 따라 로지스틱 모형으로 구한 집합 S에 포함된 v2, v10, v12, v15, v16, v17, v18을 조사한 결과, 상자도형을 보면 v18이 가장 분류능력이 좋고, 상관도를 보면 v12와 v10은 서로 연관관계가 강해 둘 중 하나만 선택되는 것이 바람직하다. 또한 상관도에서 v18은 마지막 클래스 값, v12는 두 번째 클래스 값을 명확하게 구별해 주며, v2가 추가되면 마지막과 6번째 클래스 값을 잘 구별해주는 것을 확인할 수 있다. 히스토그램보다는 4-구간 방법으로 이산화시켜 얻은 {v18, v12, v2}가 다른 이산화방법의 결과보다 더 효율적이다.



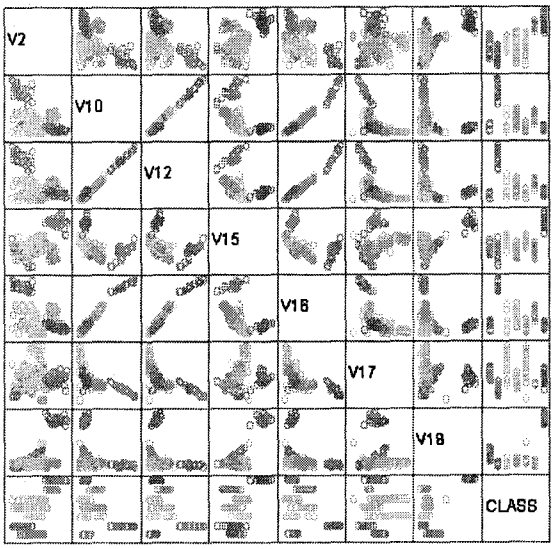
<그림 1> WINE - 상자도형



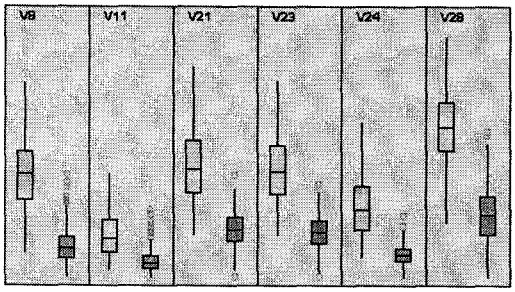
<그림 2> WINE - 상관도



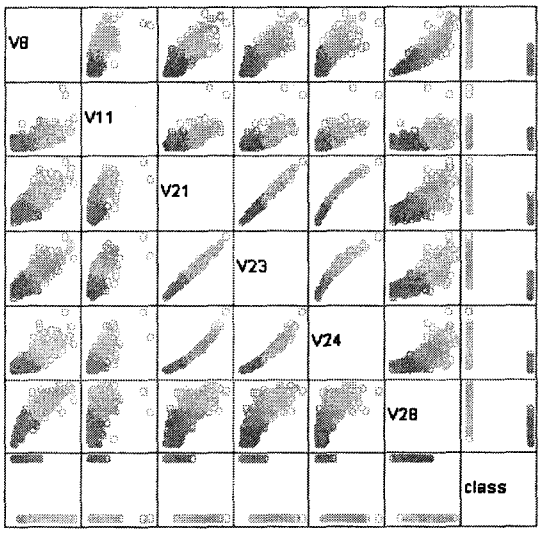
<그림 3> IMAGE - 상자도형



<그림 4> IMAGE - 상관도



<그림 5> WDBC- 상자도형



<그림 6> WDBC - 상관도

(3) WDBC 데이터베이스

마찬가지 방법으로 v8, v11, v21, v23, v24, v28에 대한 상관도를 보면, v21과 v28 두 개로 충분한 분류능력을 갖는 것을 알 수 있다. v24와 v21은 연관관계가 강해 둘 중에 하나만 선택되면 되는데 상자도형에서 보면 v21이 더 강한 분류능력을 보임을 알 수 있고 또 v11은 중요한 변수가 아님을 알 수 있다. WDBC 데이터베이스에서도 히스토그램 방법보다는 4-구간 방법으로 이산화 시켜 얻어진 {v21, v28}이 다른 이산화방법의 결과보다 더 효율적인 것을 알 수 있다.

5. 분석 및 평가

MI에 의한 변수선택방법을 연속형 데이터베이스에 적용할 때, 계산상의 문제점을 해결하기 위해서는 예측변수 값을 이산화 시킨 후 목표변수에 영향을 미치는 중요한 변수들을 선택할 수 있다. 본 논문에서는 MDL, 히스토그램, 4-구간의 세 가지 방법으로 이산화를 수행하였다. 선택된 중요변수 부분집합의 변수 수는 전체 변수에 비해 아주 적으면서도 분류 정확도는 전체변수를 다 사용한 경우보다 거의 더 좋게 나타났다. IRIS와 같은 분별력이 아주 좋은 연속형 데이터베이스에서는 이산화 방법이 변수선택에 별 영향을 미치지 않지만 데이터 구조가 복잡한 경우에는 중요한 역할을 한다. 이산화 방법으로 많이 사용하는 히스토그램 방법은 관측값의 크기에 따라 구간수가 많아지므로 계산시간이 많이 걸리고 변수마다 구간의 개수가 차이가 있을 수 있다. 또 두꺼운 꼬리를 가지거나 한쪽으로 치우쳐있는 분포의 경우 어떤 구간에는 상대적으로 많은 관측값이 속하고 어떤 구간에는 속하는 관측값이 아주 희박한 경우가 생길 수 있어서 MI를 계산하는데 문제가 된다. 또한 MDL 방법은 지도학습 방법이기 때문에 이산화를 시킬 때 예측변수 자체만 고려하는 것이 아니라 목표변수의 정보를 이용한다. 따라서 이 방법을 사용하여 변수선택을 하는 경우 새로운 종류의 데이터에 대한 분별력이 떨어질 수 있다. 또 이 방법에서도 변수마다 구간 수가 달라질 수 있다.

반면에 4-구간 방법은 모든 변수를 동일한 4개의 구간으로 나누며, 데이터가 특이값을 가지고 있거나 분포가 한쪽으로 기울어진 데이터에서도 적용할 수 있는 로버스트한 이산화방법이다. 여러 가지 실험을 해본 결과, 4-구간 방법으로 이산화시켜 MI에 의해 변수선택을 한 경우에 이산화 방법이 아주 간단함에도 불구하고 MDL, 히스토그램 방법에 비해 더 좋은 분류결과를 얻을 수 있었다. 따라서 MI를 사용하는 변수선택 문제에서 4-구간 방법을 효율적인 이산화방법으로 사용할 수 있다. 또한 이 방법은 계산시간이 적게 걸리면서도 좋은 분류결과를 얻을 수 있는 장점을 가지고 있다.

참고문헌

[1] Battiti, R.(1994). Using mutual information for selecting features in supervised neural net learning, *IEEE Transactions on Neural Networks*, 5, 537-550.

[2] Bonnlander, B. V. and Weigend, A. S.(1994). Selecting Input Variables Using mutual Information and Nonparametric Density Estimation, *Proceedings of the International Symposium on Artificial Neural Networks(ISANN)*, Taiwan, 42-50.

[3] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J.(1984). *Classification and regression trees*, Wardsworth, Belmont, CA.

[4] Cha, W. and Huh, M.(2003). Evaluation of Attribute Selection Methods and Prior Discretization in Supervised Learning, 한국통계학회 논문집, Vol. 10, No. 3, 879-894.

[5] Cover, T. M. and Thomas, J. A.(1991). *Elements of Information Theory*, Wiley, New York.

[6] Dash, M. and Liu, H.(1997). Feature selection for classification, *Intelligent Data Analysis*, Elsevier Science Inc.

[7] Devijver, P. A. and Kittler, J.(1982). *Pattern Recognition : A Statistical Approach*,

Prentice Hall International.

[8] Fayyad, U. M. and Irani, K. B.(1992). On the Handling of Continuous-valued Attributes in Decision Tree Generation, *Machine Learning*, Vol. 8, 87-192.

[9] Huh, M. Y.(2005). DAVIS(http://stat.skku.ac.kr/myhuh/DAVIS.html)

[10] Ihaka, R. and Gentleman, R.(1996). R:A language for data analysis and graphics, *Journal of Computational and Graphical Statistics*, 5(3), 299-314. (http://www.r-project.org)

[11] Liu, H. and Motoda, H.(1998). *Feature Selection for Knowledge Discovery and Data Mining*, Kluwer Academic Publishers.

[12] Merz, C. J. and Murphy, P. M.(1996). UCI Repository of Machine Learning Databases, Department of Information and Computer Science, University of California, Irvine, CA(http://www.ics.uci.edu/~mlearn/MLRepository.html)

[13] Parzen, E.(1962). On the estimation of probability density function and mode, *Annals of Mathematical Statistics*, 33(3), 1065-1076.

[14] Tourassi, G. D., Frederick, E. D., Markey, M. K. and Floyed, C. E., Jr.(2001). Application of the mutual information criterion for feature selection in computer-aided diagnosis, *Medical Physics*, 28(12), 2394-2402.

[15] Venables, W. N. and Ripley, B. D.(1994). *Modern Applied Statistics with S-Plus*, Springer, New York.

[16] Witten, I. and Frank, E.(1999). *Data Mining*, Morgan and Kaufmann. (http://www.cs.waikato.ac.nz/ml/weka)

[2005년 6월 접수, 2005년 10월 채택]