The Korean Communications
in Statistics Vol. 10 No. 3, 2003
pp. 879-894

Evaluation of Attribute Selection Methods and Prior Discretization in Supervised Learning[1)]

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

Abstract

We evaluated the efficiencies of applying attribute selection methods and prior discretization to supervised learning, modelled by C4.5 and Naive Bayes. Three databases were obtained from UCI data archive, which consisted of continuous attributes except for one decision attribute. Four methods were used for attribute selection : MDI, ReliefF, Gain Ratio and Consistency-based method. MDI and ReliefF can be used for both continuous and discrete attributes, but the other two methods can be used only for discrete attributes. Discretization was performed using the Fayyad and Irani method. To investigate the effect of noise included in the database, noises were introduced into the data sets up to the extents of 10 or 20 %, and then the data, including those either containing the noises or not, were processed through the steps of attribute selection, discretization and classification. The results of this study indicate that classification of the data based on selected attributes yields higher accuracy than in the case of classifying the full data set, and prior discretization does not lower the accuracy.

Keywords: attribute selection, discretization, classification

1. 서론

데이터베이스를 구성하는 대상(object)들은 속성(attribute, feature)과 속성의 값(value)으로 묘사되며, 속성들은 조건속성(condition attribute)과 결정속성(decision attribute)으로 이루어진다. 여기에서 사용한 용어들은 기계학습(machine learning)에서 사용하는 것으로 대상, 속성은 통계학에서의 사례(instance, case), 변수(variable)에 해당하고, 조건속성은 독립변수(independent variable) 또는 예측변수(predictor variable), 결정속성은 종속변수(dependent variable) 또는 목표변수(target variable)에 해당한다. 본 논문에서는 기계학습에서의 용어를 사용하기로 한다. 데이터베이스의 크

1) This research was financially supported by Engineering Research Center of Hansung University in the year of 2003.

2) Professor, Division of Computer Engineering, Hansung University, Seoul, 136-792, Korea
E-mail : wcha@hansung.ac.kr

3) Professor, Department of Statistics, SungkyunK\kwan University, Seoul, 110-745, Korea
E-mail : myhuh@skku.ac.kr

기는 속성들의 수와 대상들의 개수로 나타낼 수 있다. 속성들의 수가 매우 많을 경우 결정속성에 영향을 미치지 않는 부적절(irrelevant)하고 불필요한(redundant) 속성들이 다수 포함되어 있을 수 있다. 데이터베이스의 크기가 매우 클 때 부적절하고 불필요한 속성들을 미리 제거할 수 있다면 학습시간을 줄일 수 있을 뿐만 아니라 보다 명료하고 일반적인 지식을 추출할 수 있다. 중요한 속성들을 선택하는 문제는 이미 통계학 분야에서도 많이 연구되어왔으며, 패턴 인식 분야에서도 많이 다루어진 분야이다(Devijver and Kittler, 1982, Miller, 1990). 중요한 속성들을 선택하기 위한 방법 중에는 조건속성이 연속형 값을 가지는 경우에 원래 데이터를 그대로 사용하는 방법과 연속형 값을 이산화(discretizing)시킨 후에 사용할 수 있는 방법이 있다.

본 논문의 목적은 다수의 부적절한 속성이 포함되어 있는 연속형 속성들로 이루어진 데이터베이스에서 중요한 속성을 선택하는 방법들을 비교연구 하고자 하는 것이다. 비교 연구를 위하여 원래의 연속형 데이터에 바로 적용할 수 있는 속성선택방법과 연속형 데이터를 이산화 시킨 후에 적용할 수 있는 속성선택방법으로 나누고, 실제 데이터베이스와 실제 데이터베이스에 잡음(noise)을 10%, 20%씩 추가한 데이터베이스에 적용한다. 대표적인 지도학습방법인 결정나무 C4.5와 Naive Bayes 모형을 사용하여 원래 데이터베이스를 그대로 사용한 분류 결과와, 연속형 데이터에 바로 적용할 수 있는 속성선택방법의 경우 속성선택 후 분류 결과, 속성선택을 하고 이산화 시킨 후의 분류결과 등을 비교 분석한다. 이산형 데이터에 적용할 수 있는 속성선택방법의 경우는 연속형 속성 값을 먼저 이산화 후 중요한 속성을 선택하여 분류결과를 얻고, 앞의 결과와 비교 분석한다. 이를 통하여 속성선택 방법과 이산화가 분류에 어떤 영향을 주는지 알아보고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 속성선택방법들에 대해 정리하고, 연속형 데이터를 이산화 하는 방법에 대해 설명한다. 3장에서는 본 연구에서 수행하고자 하는 실험방법과 실험에 사용하고자 하는 데이터의 설명 및 데이터에 잡음을 부여하는 방법을 기술한다. 다음으로 4장에서는 실험결과와 이산화 및 속성선택이 분류에 미치는 영향을 분석하고 5장에서는 평가 및 결론을 기술하였다.

2. 속성선택방법 및 이산화방법

속성선택은 데이터베이스에서 부적절하고 불필요한 정보를 가능한 한 많이 찾아내어 제거하는 과정이다. 데이터의 차원을 축소시키면 학습 과정이 보다 빨리 효율적으로 수행될 수 있으며, 목표 개념에 대한 학습 결과가 보다 간결하고 이해하기 쉽게 된다.

속성선택 문제는 결정속성에 영향을 미치는 조건속성에 대해 중요한 속성 순으로 순위를 정하거나, 조건속성들의 집합 중 결정속성에 영향을 미치는 가장 적절한 속성들의 부분집합을 찾는 탐색 문제이며, 탐색방법과 속성들에 대한 평가방법에 따라 다양한 속성선택방법들이 연구되었다. 속성선택방법은 필터(filter)와 포장(wrapper)방법으로 분류할 수 있는데, 필터방법은 속성들을 평가할 때 학습 알고리즘과는 독립적으로 데이터가 가지고 있는 성질을 이용하는 방법이고, 포장방법은 학습 알고리즘에 의한 분류 정확도를 사용하여 속성들을 평가하는 방법이다. 필터방법에서 평가에 사용되는 척도로는 종속성(dependency), 거리(distance), 정보(information), 일관성(consistency)이 있다(Das and Liu, 1997, Liu and Motoda, 1998). 종속성은 하나의 조건속성이 결정속성에 얼마나 강하게 연관되어 있는가를 측정하는 척도이고, 거리는 결정속성의 값을 가능한 한 멀리 분리해 줄 수 있는 조건속성을 찾는 척도이며, 정보는 결정속성의 불확실성을 줄여 줄 수 있는 조건속성을

찾는 척도로 사용된다. 또, 일관성은 모든 속성이 다 사용되었을 때와 동일하게 결정속성의 값을 분류할 수 있는 최소의 조건속성을 찾는데 사용할 수 있는 척도이다. 속성선택방법들의 비교평가를 위한 기존 연구로는 Dash and Liu(1997), Hall and Homes(2001)의 연구 등이 있다.

2.1 속성선택방법

본 논문에서 비교 평가하고자 하는 속성선택방법은 종속성척도를 사용하는 MDI(Lee and Huh, 2003), 거리척도를 사용하는 ReliefF(Kononenko, 1994), 정보척도를 사용하는 정보획득비(Gain Ratio, Quinlan, 1998), 그리고 일관성척도를 사용하는 일관성기반방법(Consistency-based, Liu and Setino, 1996)이다. MDI와 ReliefF는 연속형과 이산형 속성이 혼합되어있는 경우에 바로 적용할 수 있는 방법이며, 정보획득비와 일관성기반방법은 이산형 속성에만 적용할 수 있는 방법이다. 또한 MDI, ReliefF, 정보획득비 방법은 중요한 속성 순으로 순위를 정해주는 방법이고, 일관성기반방법은 조건속성들의 집합 중 결정속성에 영향을 미치는 가장 적절한 속성들의 부분집합을 구해주는 방법이다. 각 방법에 대해 간략하게 설명한다.

(1) MDI방법

MDI(Measure of Departure from Independence) 방법에서는 데이터베이스를 구성하고 있는 속성들이 연속형, 이산형 일 때, 두 속성간의 연관성 측정을 위해 두 속성간의 독립성 검정방법을 사용한다. 두 속성이 모두 연속형인 경우 두 속성간의 독립성 검정은 Spearman의 순위상관계수 검정을 사용하고 있으며, 연속형과 이산형인 경우는 Kruskal-Wallis 검정을, 그리고 두 속성 모두 이산형인 경우 Pearson의 카이제곱검정을 사용하고 있다. 이들 검정통계량을 사용하여 두 속성간의 독립성 검정을 수행할 때 나타나는 유의확률 p를 이용하여 속성들 간의 연관성을 측정하는 것이 MDI 방법이다.

(2) ReliefF방법(REL)

Relief방법(Kira and Rendell, 1992)은 속성들이 서로 독립일 때와, 속성들이 서로 강한 종속성(strong dependency)을 가지고 있을 때에도 속성들의 중요도를 추정하여 결정속성에 영향을 미치는 속성들의 순위를 결정할 수 있는 방법이다. 이 방법도 MDI와 마찬가지로 속성들이 연속형, 이산형 값을 가지는 데이터베이스에 적용할 수 있다. Relief방법에서는 데이터베이스에서 하나의 대상을 임의로 추출하여, 이 대상과 결정속성 값이 같은 클래스와 다른 클래스로부터 최근접 이웃을 찾아내고 이 대상의 속성들의 값과 각 클래스에 속하는 최근접 이웃들에 속하는 속성들의 값을 비교하여 각각의 속성들의 점수를 갱신하는데 사용한다. 이 과정을 사용자가 미리 정한 대상들의 수가 될 때까지 반복하여 결정속성에 영향을 미치는 속성들의 순위를 정한다. 만약 어떤 속성이 유용하다면 서로 다른 클래스에 속하는 대상들에서는 다른 값을 가지고, 같은 클래스에 속하는 대상들에서는 같은 값을 가져야 한다. Relief는 이러한 원리에 바탕을 둔 것이다. Relief는 결정속성이 가지는 값이 두 개일 때 사용할 수 있는 방법이며, 데이터가 결측값을 많이 포함하고 있거나, 결정속성 값이 세 개 이상인 경우에 사용할 수 있도록 확장한 방법이 ReliefF이다. 이 방법에서는 추출된 대상과 같은 클래스 또는 다른 클래스로부터 하나의 최근법 이웃을 찾는 대신 k개의 최근접 이웃(k-nearest neighbor)을 찾아 평균을 구함으로서 데이터에 내재하는 잡음의 영향을 줄여준다. 세 개 이상의 결정속성 값을 가지는 데이터의 경우 각각의 클래스에 대한 최근접 이웃을 찾고 각각의 클래스의 사전 확률 값으로 가중치를 부여하여 속성들의 점수를 계산한다.

(3) 정보획득비방법(RATIO)

C 와 D 를 조건속성과 결정속성이라 할 때, 조건속성을 관찰하기 전과 후의 결정속성의 엔트로피는 다음 식과 같다.

$$H(D) = -\sum_{d \in D} p(d) \log(p(d))$$

$$H(D|C) = -\sum_{c \in C} p(c) \sum_{d \in D} p(d|c) \log(p(d|c))$$

결정속성의 엔트로피가 감소하는 양은 조건속성에 의해 제공되는 결정속성에 대한 추가적인 정보를 반영하는 것이고 각각의 조건속성 C_i 에 대한 정보획득(Information gain)은 $H(D) - H(D|C_i)$ 와 같다(Quinlan, 1986). 정보획득은 결정속성이 갖는 값이 많아질 때 값이 커지는 경향이 있어 이를 보완하기 위해 일종의 표준화를 시도하며 이를 정보획득비(Quinlan, 1998)라 하고 다음과 같이 정의한다.

$$GainR(D, C_i) = H(D) - H(D|C_i) \ / \ H(C_i)$$

이 방법에서는 각각의 조건속성과 결정속성 사이의 정보획득비에 의해 조건속성들의 순위가 결정된다.

(4) 일관성기반방법(CON)

속성들의 부분집합이 결정속성에 미치는 영향을 평가하는 방법으로서 데이터베이스를 구성하는 모든 속성들을 다 사용하였을 때와 동일한 일관성을 가지는 가장 작은 속성들의 부분집합을 탐색하는 방법이다. 일관성 척도는 다음과 같이 구한다(Liu and Setino, 1996).

우선 두 대상이 결정속성 값만 다를 때 비일관성(inconsistency)을 가진다고 정의한다. 결정속성 값은 고려하지 않고 모든 일치하는 대상들에 대해 비일관성 점수(inconsistency count)는 대상들의 수에서 결정속성 값이 동일한 대상들의 수 중 가장 큰 값을 뺀 값이다. 예를 들어, n 개의 일치하는 대상들이 있고, $c1$ 개는 결정속성 D_1 값을 가지고, $c2$ 개는 D_2 , $c3$ 개는 D_3 값을 가지고 (여기서, $c1 + c2 + c3 = n$), $c3$ 가 가장 큰 값이면 비일관성 점수는 $n - c3$ 이다. 비일관성 비율 (inconsistency rate)은 모든 비일관성 점수의 합을 전체 대상들의 수로 나눈 값이고, 일관성 척도는 1에서 비일관성 비율을 뺀 값이다. 이상을 수식으로 정리하면 다음과 같다.

$$CON_s = 1 - \frac{\sum_{i=0}^{J} (|D_i| - |M_i|)}{N}$$

여기에서, s 는 속성들의 부분집합이고, J 는 s 에 속하는 속성들의 속성 값들의 조합의 수, $|D_i|$ 는 i 번 째 속성 값 조합의 발생빈도수, $|M_i|$ 는 i 번 째 속성 값 조합에 대해서 결정속성 값이 동일한 대상들의 수 중 가장 큰 값이다. N 은 데이터베이스의 대상들의 총 수이다.

2.2 이산화방법

이산형 속성에만 적용할 수 있는 속성선택방법을 연속형 속성에 사용하기 위해서는 먼저 연속형 속성 값을 이산형 속성 값으로 변환하는 과정이 필요하다. 또 결정나무나 연관성 규칙을 찾아내는 것과 같은 많은 데이터마이닝 도구들의 알고리즘은 이산형 속성 값을 갖는 데이터에 기초를 두고 있다. 연속형 속성을 포함하는 데이터에 대한 이산화 방법은 전역적 이산화(global discretization)와 지역적 이산화(local discretizatiopn)로 나눌 수 있는데, 전역적 이산화는 학습 알고리즘을 사용하기 이전에 전체 데이터에 대해 수행하며, 지역적 이산화는 결정나무의 생성 과정에서와 같이 각 노드에서 그 노드에 도달한 데이터만을 이용하여 부분적으로 이산화를 수행한다. 전역적 이산화 방법은 지도학습의 전처리 과정으로 유용하게 사용할 수 있으며 대표적인 방법은 Gini Index(Breiman et al, 1984), 정보획득 등 불순도 척도(impurity measure)에 바탕을 둔 방법이다. 본 연구에서는 연속형 속성들 중에서 중요한 속성을 선택을 위한 전역적 이산화와 지도학습 방법을 적용하기 위한 전역적 이산화 방법으로서 Fayyad와 Irani(1992)의 MDL 방법을 사용한다.

이 방법은 정보획득에 바탕을 둔 이산화 방법으로서 분할점(split point)을 선택하기 위해 후보 분할점(candidate split point)들의 평균 클래스 엔트로피(average class entropy) $E(C,T;S)$ 를 이용한다.

S 를 데이터집합, C 를 연속형 속성, T 를 후보 분할점, 그리고 S_1, S_2 를 S 를 T 로 분할한 경우의 부분구간이라 할 때 S 의 클래스 엔트로피 함수와 평균 클래스 엔트로피는 다음과 같다.

$Ent(S) = -\sum_{i=1}^{k} P(D_i, S)\log(P(D_i, S))$, 여기서 $P(D_i, S)$ 는 S 중에서 클래스 D_i 에 속할 확률이다.

$E(C,T;S) = \frac{|S_1|}{N} Ent(S_1) + \frac{|S_2|}{N} Ent(S_2)$, 여기서 $N = |S|$ 이다.

이 방법에서는 다중 분할점을 생성하기 위해 연속형 속성 C 에 대해서 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 이진분할 하는 방법을 사용하였다. 그리고 이렇게 진행하는 이진분할은 다음과 같은 조건을 가질 때 중지하게 된다.

$$\text{정보획득} > \frac{\log(N-1)}{N} + \frac{\log(3^k-2) - kE + k_1E_1 + k_2E_2}{N}$$

여기서 log 는 밑을 2 로 하며, $E = Ent(S)$, $E_1 = Ent(S_1)$, $E_2 = Ent(S_2)$ 이고 k_1, k_2 는 각 분할에 속해 있는 클래스의 수이다.

3. 실험

본 연구를 위한 실험설계에서는 다음과 같은 6가지 요인들을 고려하였다.

1. 지도학습방법: C4.5, Naive Bayes
2. 속성선택방법: MDI, ReliefF, 정보획득비, 일관성기반 방법
3. 데이터: UCI 데이터 창고의 Ionosphere, Wine Recognition, Pima Indians Diabetes
4. 잡음의 영향: 10%, 20%의 잡음을 데이터에 부과
5. 속성을 선택하여 지도학습을 하는 방법과 속성 전체를 지도학습에 적용시키는 방법.
6. 연속형 속성을 이산화 시키고 지도학습에 적용하는 방법(사전 이산화)과 연속형 속성을 그대로 지도학습에 적용시키는 방법.

이들 요인들을 고려하여 실험설계를 하고, 각 실험은 10-층 교차타당성(10-fold cross validation) 실험을 10번 수행한 분류결과를 평균하여 비교 평가한다. 속성선택방법 중에서 MDI 방법은 R(Ihaka and Gentleman, 1967)을 사용하였으며, 나머지 방법들은 공개 소프트웨어 WEKA(Witten and Frank, 1999)를 사용하였다. WEKA *Explorer*에서 속성선택방법을 사용할 때, Attribute Evaluator로서 ReliefF 방법은 *ReliefFAttributeEval*, 정보획득비방법은 *GainRatioAttributeEval*, 일관성기반방법은 *ConsistencySubsetEval*을 사용하였으며, 탐색방법(Search Method)은 ReliefF방법과 정보획득비방법에서는 *Ranker*, 일관성기반방법에서는 *Best First*를 사용하였다. 이산화를 위해서는 R 프로그램과 WEKA의 *DiscretizeFilter* 프로그램을 수정하여 사용하였다. 분류결과는 WEKA의 *Experimenter*에서 Destination은 *InstanceResultListener*, Result generator는 *AveragingResultProducer*를 사용하여 얻었으며, 분류모형은 WEKA의 classifier *j48*, *NaiveBayes*를 사용하였고 기본 사양을 적용하여 실험하였다.

실험과정을 구체적으로 기술하면 다음과 같다.

3.1 {속성선택, 이산화, 분류} 의 순서

1. MDI, ReliefF 방법을 적용하여 속성선택을 하는 경우

이 방법의 경우, {속성선택, 이산화, 분류}의 순서에 따라 다음과 같은 조합에 대해 실험하였다.

(1) **속성선택→분류** : 원래의 연속형 데이터에 대해 MDI와 ReliefF방법을 사용하여 중요한 속성을 선택한 후, 분류모형을 적용시킨다.

(2) **속성선택→이산화→분류** : 원래의 데이터에 대해 중요 속성을 선택한 후, 데이터를 이산화 시키고 분류모형을 적용시킨다.

(3) **이산화→속성선택→분류** : 연속형 데이터에 그대로 사용할 수 있는 방법이지만 이산화의 영향을 알아보기 위해 먼저 이산화 시킨 후 중요한 속성을 선택하고 분류모형을 적용시킨다.

2. 정보획득비, 일관성기반 방법을 적용하여 속성선택을 하는 경우

이 방법들은 이산형 데이터에만 적용할 수 있으므로 다음과 같은 순서에 따라 실험하였다.

이산화→속성선택→분류

즉, 데이터를 사전 이산화 시키고, 두 방법에 따라 중요한 속성을 선택하고 분류모형을 적용시킨다.

3.2 중요한 속성선택의 기준

{ $C_1, C_2, \ldots, C_p$ }를 p 개의 조건속성이라 하고 D 를 결정속성이라 하자. 데이터베이스에 MDI, ReliefF, 정보획득비 속성선택방법을 적용시킨다. 각 방법에 의해 중요한 속성으로 구해진 조건속성들을 결정속성에 영향을 많이 미치는 순(중요도 순)으로 순서배열하고 이를 $C_{(1)}, C_{(2)}, \ldots, C_{(p)}$ 라고 하자. 이제 i 개의 조건속성과 결정속성으로 이루어진 데이터 파일 $S_{(i)}$ 를 다음과 같이 만든다. $S_{(i)}$ = { $C_{(1)}, C_{(2)}, \ldots, C_{(i)}, D$ }. 또한 일관성기반 방법으로 구한 중요한 조건속성들의 부분집합과 결정속성으로 이루어진 데이터 파일도 생성한다. 이제 $S_{(i)}$ ($i=1, \ldots, p$)에 대해 WEKA의 *Experimenter*를 사용하여 C4.5와 Naive Bayes 모형에 대한 10-층 교차타당성 실험을 10번 수행한다. 분류결과 각 $S_{(i)}$ 에 해당하는 정확도 (accuracy) $A_{(i)}$ 를 구하면, $A_{(i)}$ 는 i 에 대해 증가하다가 어느 시점에서 감소하게 되는데 감소하기 직전 데이터 파일의 조건속성들을 중요한 속성으로 선택한다.

3.3 사용 데이터베이스

본 논문의 실험에 사용한 데이터베이스는 UCI 데이터 창고(Merz and Murphy, 1996)에 있는 Ionosphere, Wine Recognition, Pima Indians Diabetes로서 조건속성들은 모두 연속형 값을 갖는다.

(1) Ionosphere 데이터베이스(IONO)
데이터베이스의 크기는 351이고 2개의 값을 갖는 결정속성(good(225), bad(126))과 34개의 연속형 조건속성으로 이루어져 있으며 결측값은 없다. 잡음을 각각 10%, 20%씩 부여하여 IONO+10, IONO+20 데이터베이스를 생성한다.

(2) Wine Recognition 데이터베이스(WINE)
데이터베이스의 크기는 178이고 3개의 값을 갖는 결정속성(1(59), 2(71), 3(48))과 13개의 연속형 조건속성으로 이루어져 있으며 결측값은 없다. 잡음을 각각 10%, 20%씩 부여하여 WINE+10, WINE+20 데이터베이스를 생성한다.

(3) Pima Indians Diabetes 데이터베이스(PIMA)
데이터베이스의 크기는 768이고 2개의 값을 갖는 결정속성(0(500), 1(268))과 8개의 연속형 조건속성으로 이루어져 있으며 결측값은 없다. 잡음을 각각 10%, 20%씩 부여하여 PIMA+10, PIMA+20 데이터베이스를 생성한다.

3.4. 잡음부여방법

실험에 사용하는 데이터베이스에 대해 전체 대상 중 무작위로 10%, 20%의 대상을 선택하여 결정속성 값을 다른 값들 중에서 무작위로 골라 부여한다.

4. 실험결과 및 분석

4.1. 속성선택을 한 후 분류하는 경우 MDI와 ReliefF 방법의 효율비교

MDI와 ReliefF(REL)는 혼합형 데이터에 적용할 수 있는 속성선택방법이므로, 여기서는 이 두 가지 방법의 효율을 비교한다.

4.1.1 속성선택 결과의 비교

MDI와 ReliefF방법을 적용하여 IONO, WINE, PIMA 데이터와 여기에 10%, 20%의 잡음을 첨가한 데이터에서, 중요도에 따라 선택한 조건속성 순위는 다음 [표 1], [표 2], [표 3]과 같다. 여기서 편의를 위해 IONO와 WINE 데이터의 경우 처음 10개 순위에 해당하는 속성번호만 기재하였다.

[표 1] IONO 데이터에서 속성선택 결과 중요도에 따른 속성 번호

	IONO	IONO + 10	IONO + 20
MDI	1, 3, 7, 5, 27, 33, 31, 9, 8, 15	1, 27, 3, 7, 5, 9, 31, 8, 14, 33	1, 27, 3, 7, 5, 9, 31, 8, 14,33
REL	24, 3, 8, 5, 14, 7, 16, 34, 29, 9	3, 5, 27, 24, 8, 15, 33, 11, 12, 22	3, 5, 27, 24, 8, 15, 33, 11, 12, 22

[표 2] WINE 데이터에서 속성선택 결과 중요도에 따른 속성 번호

	WINE	WINE + 10	WINE + 20
MDI	7, 13, 10, 1, 12, 6, 11, 4, 9, 2	7, 12, 10, 13, 1, 6, 11, 4, 2, 8	7, 13, 6, 10, 12, 11, 1, 9, 4, 2
REL	12, 7, 13, 1, 10, 6, 11, 8, 2, 9	13, 10, 1, 12, 7, 8, 6, 2, 11, 3	10, 13, 1, 7, 12, 6, 8, 3, 11, 9

[표 3] PIMA 데이터에서 속성선택 결과 중요도에 따른 속성 번호

	PIMA	PIMA + 10	PIMA + 20
MDI	2, 6, 8, 1, 7, 3, 4, 5	2, 6, 8, 1, 7, 3, 4, 5	2, 6, 8, 1, 7, 3, 5, 4
REL	2, 6, 4, 1, 8, 7 ,3, 5	2, 4, 6, 1, 8, 3, 5, 7	2, 1, 6, 4, 3, 8, 7, 5

이 실험결과를 보면 두 속성선택방법에 의한 중요한 속성들의 순위는 방법에 따라 약간의 차이가 있다. 잡음을 첨가한 데이터에서 선택된 속성들이 원래 데이터에서 선택된 속성과 같은 경우는 진하게 표시하였는데, IONO 데이터에 대한 ReliefF방법을 제외하면 중요도에 따른 순서는 조금 다르지만 선택된 속성들은 거의 차이가 나지 않았다. 따라서 두 가지 속성선택방법은 비교적 잡음에 강한 것을 알 수 있다.

4.1.2. 속성선택한 후 이를 사용하여 분류하는 경우 정확성 비교

원래 데이터에 주어진 모든 조건속성을 다 사용하는 경우(FULL)와 MDI 및 ReliefF방법으로 3.2절에서 정한 기준에 따라 선택된 조건속성만 사용하는 경우에 대해 C4.5와 Naive Bayes 모형에 적용하여 분류한다. 비교 기준은 10-층 교차타당성 실험을 10번 수행하여 구한 분류결과의 정확도(%)이다. 결과는 [표 4], [표 5], [표 6]과 같다. 괄호 안은 선택된 조건속성들의 개수와 전체 조건

속성에 대한 백분율을 나타낸다. FULL에 해당하는 행에는 10-층 교차타당성 실험을 10번 수행하여 구한 분류정확도의 평균(%)을 나타내었고, 나머지 행에 주어진 자료는 FULL에 대한 상대 정확도이다. 예를 들어 IONO +10 데이터의 경우, C4.5 분류방법에 대한 MDI에서의 정확도는 81.59%이며 이는 FULL에 대해 103%의 상대 정확도임을 나타내고 있다.

[표 4] IONO 데이터에 대한 분류정확도

	IONO		IONO +10		IONO +20	
	C4.5	NB	C4.5	NB	C4.5	NB
FULL	89.74	82.17	79.26	75.61	65.92	69.88
MDI	102 (5/14.7%)	107 (5/14.7%)	103 (3/8.8%)	105 (3/8.8%)	109 (3/8.8%)	102 (3/8.8%)
REL	103 (4/11.8%)	110 (5/14.7%)	103 (4/11.8%)	109 (7/20.6%)	107 (3/8.8%)	104 (5/14.7%)

[표 5] WINE 데이터에 대한 분류 정확도

	WINE		WINE +10		WINE +20	
	C4.5	NB	C4.5	NB	C4.5	NB
FULL	80.26	85.61	64.39	73.67	58.85	69.84
MDI	119 (3/23.1%)	112 (4/30.8%)	112 (4/30.8%)	102 (4/30.8%)	115 (3/23.1%)	96 (2/15.4%)
REL	113 (3/23.1%)	112 (7/53.8%)	114 (4/30.8%)	103 (4/30.8%)	108 (4/30.8%)	102 (4/30.8%)

[표 6] PIMA 데이터에 대한 분류 정확도

	PIMA		PIMA +10		PIMA +20	
	C4.5	NB	C4.5	NB	C4.5	NB
FULL	74.49	75.75	67.73	69.12	65.84	69.14
MDI	100 (2/25%)	101 (2/25%)	101 (4/50%)	102 (3/37.5%)	103 (3/37.5%)	102 (2/25%)
REL	100 (2/25%)	101 (2/25%)	99 (1/12.5%)	101 (4/50%)	102 (4/50%)	101 (3/37.5%)

이 결과를 보면 MDI와 ReliefF 방법에 의해 선택된 속성만을 사용하여 분류를 하더라도 대부분의 경우(36번 중 34번) 모든 속성을 다 사용한 경우보다 정확도가 좋아진 것을 알 수 있다. 특히 IONO 데이터의 경우 전체 34개의 조건 속성 중에서 3개 ~ 7개의 조건속성만 사용하더라도 분류결과 정확도는 매우 좋은 것을 알 수 있다. WINE, PIMA 데이터의 경우도 원래 조건속성 수의 50% 미만을 사용하여 좋은 결과를 얻었다. 두 가지 속성선택방법 중 MDI 방법에 의한 결과가 ReliefF 방법에 의한 결과보다 더 좋은 경우가 전체 18번의 분류결과 중 7번, ReliefF의 결과가 더 좋은 경우가 7번이고, 같은 경우가 4번이다.

4.1.3. 속성선택한 후 이산화를 하고 분류하는 경우 정확성 비교

원래 데이터에 주어진 모든 조건속성을 다 사용하는 경우(FULL)와 MDI 및 ReliefF방법에 의해 선택된 조건속성만 사용하고 선택된 속성 값들을 Fayyad와 Irani 방법으로 이산화 시킨다. 이산화된 데이터를 사용하여 C4.5와 Naive Bayes 모형에 적용하여 분류한다. 비교 기준은 앞에서와 마찬가지로 10-층 교차타당성 실험 10번을 수행하여 구한 분류결과의 정확도(%)이다. 결과는 [표 7], [표 8], [표 9]와 같다. 괄호 안은 선택된 조건속성들의 개수와 전체 조건속성에 대한 백분율을 나타낸다.

[표 7] IONO 데이터베이스에 대한 분류정확도

	IONO		IONO + 10		IONO + 20	
	C4.5	NB	C4.5	NB	C4.5	NB
FULL	90.2	89.29	81.79	81.96	70.82	73.87
MDI	100 (3/8.8%)	103 (5/14.7%)	100 (3/8.8%)	100 (3/8.8%)	102 (3/8.8%)	98 (3/8.8%)
REL	103 (6/17.6%)	104 (4/11.7%)	101 (8/23.5%)	100 (3/8.8%)	100 (3/8.8%)	96 (3/8.8%)

[표 8] WINE 데이터베이스에 대한 분류정확도

	WINE		WINE + 10		WINE + 20	
	C4.5	NB	C4.5	NB	C4.5	NB
FULL	90.9	98.71	72.71	75.24	66.08	70.27
MDI	105 (3/23.1%)	98 (4/30.8%)	101 (5/38.5%)	101 (5/38.5%)	105 (2/15.4%)	94 (2/15.4%)
REL	102 (4/30.1%)	99 (7/53.8%)	103 (4/30.8%)	103 (4/30.8%)	104 (4/30.8%)	101 (4/30.8%)

[표 9] PIMA 데이터베이스에 대한 분류정확도

	PIMA		PIMA + 10		PIMA + 20	
	C4.5	NB	C4.5	NB	C4.5	NB
FULL	74.32	75.26	66.98	66.18	67.77	66.6
MDI	101 (4/50%)	101 (4/50%)	101 (3/37.5%)	103 (3/37.5%)	100 (3/37.5%)	101 (3/37.5%)
REL	101 (5/62.5)	100 (3/37.5%)	100 (5/62.5%)	102 (3/37.5%)	100 (5/62.5%)	101 (5/62.5%)

이 실험에서는 MDI, ReliefF방법으로 구한 데이터베이스의 전체 조건속성의 일부만 가지고 분류를 하는 경우 모든 속성을 다 사용한 경우보다 전체 36번 중 31번 정확도가 같거나 더 좋아졌다. MDI 방법에 의한 결과가 ReliefF에 의한 결과보다 더 좋은 경우가 전체 18번의 분류결과 중 7번, ReliefF의 결과가 더 좋은 경우가 7번이고 동일한 경우가 4번이다.

4.2. 이산화를 하고 속성선택을 한 후 분류하는 경우 4 가지 방법의 비교

혼합형 데이터에서 연속형 속성을 이산화 시키면 모든 속성이 이산형이 된다. 모든 속성이 이산형이면, 본 논문에서 고려하는 네 가지 속성선택방법(MDI, ReliefF, 정보획득비, 일관성기반방법)이 모두 적용 가능하므로, 여기서는 이들 방법 모두를 비교 평가할 수 있다.

4.2.1 속성선택 결과의 비교

원래 데이터를 먼저 Fayyad와 Irani 방법에 의해 이산화를 시킨다. 이렇게 이산화 시킨 데이터에 MDI, ReliefF, 정보획득비(RATIO), 일관성기반방법(CON)의 4가지 속성선택방법을 적용하여 중요한 조건속성을 선택한다. MDI, ReliefF, 정보획득비 방법에 의한 속성들의 순위와 일관성기반 방법에 의한 속성집합은 다음 [표 10], [표 11], [표 12]와 같다.

[표 10] IONO 데이터베이스의 속성들의 순위

	IONO	IONO +10	IONO +20
MDI	1, 3, 5, 27, 9, 7, 33, 31, 14, 8	**3, 7, 27, 1, 14**, 16, **9, 5, 31, 33**	**27, 7, 3, 5, 9**, 24, 22, **1**, 6, **8, 14**
REL	5, 3, 34, 8, 22, 6, 33, 7, 16, 27	**34**, 17, **5, 8**, 12, 23, 4, 21, 9, **33**	4, 12, **3, 27, 8, 34**, 20, 28, **16**, 30
RATIO	1, 28, 18, 5, 7, 20, 24, 33, 6, 27	3, **27, 1, 7**, 16, 14, **24, 28, 18, 6**	**24**, 3, **6**, 8, **7**, 34, 10, 22, **27, 18**
CON	{5, 6, 8, 13, 22, 27, 34}	{3, 4, 10, 12, 14, 17, 21, 23, 29, 33, **34**}	{1, 3, 4, **6, 8**, 10, 12, 14, 16, **22**, 24, **27**, 28, 31}

[표 11] WINE 데이터베이스의 속성들의 순위

	WINE	WINE +10	WINE +20
MDI	7, 12, 13, 10, 1, 6, 11, 4, 9, 2	**7, 10, 12, 13, 1, 6, 11, 4, 2, 9**	**7, 12, 6, 10, 13, 1, 11, 2, 9, 4**
REL	7, 10, 12, 6, 13, 2, 4, 11, 1, 8	**7, 10, 12, 4, 1, 13, 8, 2, 6**, 5	**10, 7, 11, 6, 1, 12, 2, 4, 13**, 9
RATIO	12, 7, 10, 13, 1, 6, 11, 3, 5, 4	**10, 12, 7, 13, 6**, 1, **11**, 2, **4**, 9	**7, 6, 12, 13, 10**, 11, 1, 9, 2, **5**
CON	{1, 3, 4, 7, 10}	{1, 2, **4**, 5, 6, **7**, 8, 9, **10**, 11, 12, 13}	{1, 2, **4**, 5 ,6, **7**, 8, **10**, 11, 13}

[표 12] PIMA 데이터베이스의 속성들의 순위

	PIMA	PIMA +10	PIMA +20
MDI	2, 8, 6, 1, 7, 5	**2, 6, 8, 1, 5**	**2, 6, 8, 1, 5**
REL	2, 6, 5, 7, 8, 1	**2, 5, 6, 8, 1**	**2, 8, 6, 5, 1**
RATIO	2, 6, 8, 1, 5, 7, 4, 3	**2, 6, 8, 1, 5**	**2, 6, 8, 1, 5**
CON	{1, 2, 3, 4, 5, 6, 7, 8}	{**1, 2, 5, 6, 8**}	{**1, 2, 5, 6, 8**}

이 실험결과를 보면 속성선택방법에 의한 중요한 속성들의 순위는 방법에 따라 약간의 차이가 있다. 잡음을 첨가한 데이터에서 선택된 속성들이 원래 데이터에서 선택된 속성과 같을 경우는 진하게 표시하였는데, 데이터에 잡음을 첨가하였을 때 각 방법에 의한 속성선택 결과는 많은 차이가 나지 않았다. 따라서 먼저 이산화를 하고 속성선택을 하는 경우도 비교적 잡음에 강한 것을 알 수

있다.

4.2.2. 이산화를 하고 속성선택한 후 분류하는 경우 정확성 비교

선택된 속성들로 이루어진 데이터에 C4.5와 Naive Bayes 모형을 적용하여 분류한다. 비교 기준은 앞에서와 마찬가지로 10-층 교차타당성 실험 10번을 수행하여 구한 분류결과의 정확도(%)이다. 결과는 [표 13], [표 14], [표 15]와 같고 괄호 안은 선택된 조건속성들의 개수와 전체 조건속성에 대한 백분율을 나타낸다.

[표 13] IONO 데이터베이스에 대한 분류 정확도

	IONO		IONO+10		IONO+20	
	C4.5	NB	C4.5	NB	C4.5	NB
FULL	89.25	86.92	81.17	85.24	73.53	77.52
MDI	103 (4/11.8%)	108 (7/20.6%)	102 (3/8.8%)	96 (3/8.8%)	100 (3/8.8%)	95 (3/8.8%)
REL	104 (4/11.8%)	108 (3/8.8%)	97 (1/2.9%)	94 (1/2.9%)	104 (7/20.6%)	98 (7/20.6%)
RATIO	101 (4/11.8%)	106 (7/20.6%)	103 (6/17.6%)	98 (3/8.8%)	103 (4/11.8%)	98 (4/11.8%)
CON	102 (7/20.6%)	104 (7/20.6%)	103 (11/32.4%)	99 (11/32.4%)	100 (14/41.2%)	96 (14/41.2%)

[표 14] WINE 데이터베이스에 대한 분류정확도

	WINE		WINE +10		WINE +20	
	C4.5	NB	C4.5	NB	C4.5	NB
FULL	85.45	88.36	75.56	76.97	66.77	72.23
MDI	101 (3/23.1%)	99 (6/46.2%)	102 (4/30.8%)	103 (5/38.5%)	103 (6/46.2%)	96 (3/23.1%)
REL	100 (6/46.2%)	93 (3/23.1%)	101 (6/46.2%)	103 (6/46.2%)	102 (5/38.5%)	99 (6/46.2%)
RATIO	99 (4/30.8%)	99 (6/46.1%)	100 (2/15.4%)	99 (2/15.4%)	102 (2/15.4%)	102 (6/46.1%)
CON	97 (5/38.5%)	97 (5/38.5%)	100 (12/92.3%)	100 (12/92.3%)	102 (10/76.9%)	100 (10/76.9%)

이 실험에서도 MDI, ReliefF 방법에서는 데이터베이스의 전체 조건속성의 일부만 가지고 분류를 하는 경우 모든 속성을 다 사용한 경우보다 전체 36번 중 24번 정확도가 같거나 더 좋아졌다. MDI 방법에 의한 결과가 ReliefF 방법에 의한 결과보다 더 좋은 경우가 전체 18번의 분류결과 중 9번, ReliefF의 결과가 더 좋은 경우가 4번이고, 동일한 경우가 5번이다.

따라서 4.1.2, 4.1.3의 실험결과와 4.2.2의 MDI, ReliefF 실험결과를 종합하면 중요한 속성선택을 한 후 분류를 수행하면 모든 속성을 다 사용하는 경우보다 대체적으로 좋은 결과를 얻을 수 있었고(전체 108번 중 89번), MDI의 결과가 더 좋은 경우는 전체 54번 중 23번, ReliefF의 경우는 18

번, 동일한 경우가 13번 있었다. MDI와 ReliefF 속성평가방법은 속성선택→분류, 속성선택→이산화→분류에서는 차이가 없었으며, 이산화-속성선택-분류에서는 MDI방법이 ReliefF보다 좋은 결과를 얻었다.

[표 15] PIMA 데이터베이스에 대한 분류정확도

	PIMA		PIMA +10		PIMA +20	
	C4.5	NB	C4.5	NB	C4.5	NB
FULL	77.28	77.85	67.25	69.57	70.64	70.65
MDI	102 (5/62.5%)	101 (5/62.5%)	103 (1/12.5%)	101 (3/37.5%)	100 (3/37.5%)	102 (4/50%)
REL	99 (2/25%)	98 (2/25%)	103 (1/12.5%)	99 (1/12.5%)	100 (3/37.5%)	102 (4/50%)
RATIO	101 (4/50%)	101 (4/50%)	103 (1/12.5%)	101 (3/37.5%)	101 (2/25%)	101 (2/25%)
CON	100 (8/100%)	100 (8/100%)	100 (5/62.5%)	100 (5/62.5%)	100 (5/62.5%)	100 (5/62.5%)

또 C4.5의 경우 MDI, ReliefF방법에서 원래 데이터에 이산화를 하지 않고 속성선택→분류를 수행한 것보다는 속성선택→이산화→분류, 이산화→속성선택→분류의 결과가 더 좋은 것을 알 수 있고, 이산화를 속성선택 이전에 하는 것과 이후에 하는 것에는 큰 차이가 없는 것을 확인할 수 있다. Naive Bayes의 경우도 MDI, ReliefF 방법에서 원래 데이터에 이산화를 하지 않고 속성선택→분류를 수행한 것보다는 속성선택→이산화→분류, 이산화→속성선택→분류의 결과가 더 좋은 것을 확인할 수 있고, 이 경우에는 이산화를 먼저 하는 것이 더 좋은 결과를 얻을 수 있음을 확인할 수 있었다.

정보획득비, 일관성기반 방법에서는 데이터베이스의 전체 조건속성의 일부만 가지고 분류를 하는 경우 모든 속성을 다 사용한 경우보다 전체 36번 중 27번 정확도가 같거나 더 좋아졌다. 정보획득비 방법에 의한 결과가 일관성기반 방법에 의한 결과보다 더 좋은 경우가 전체 18번의 분류결과 중 12번, 일관성기반 방법의 결과가 더 좋은 경우가 3번이고 동일한 경우가 3번이다. 따라서, 일관성기반 방법은 정보획득비 방법보다 분류에 사용하는 속성들의 수는 훨씬 많으면서도 정확도는 떨어지는 것을 알 수 있다.

이산화→속성선택→분류 실험에서 MDI, ReliefF, 정보획득비방법을 비교하기 위하여 각각의 데이터베이스에 대한 분류결과 정확도를 다음 [그림 1], [그림 2], [그림 3]으로 나타낸 결과, MDI와 정보획득비방법이 ReliefF 보다 약간 좋은 결과를 보이는 것을 알 수 있다.

따라서 앞의 결과들을 종합해 보면, 연속형 속성 값을 가지는 데이터를 먼저 이산화 시킨 후 MDI나 정보획득비 속성선택방법을 사용하여 중요한 속성을 선택하고, 분류를 수행하면 더 좋은 결과를 얻을 수 있다는 것을 알 수 있다.

5. 평가 및 결론

데이터의 양이 방대한 경우, 데이터 관리뿐만 아니라 이를 분석하는 데도 많은 문제가 나타날 수 있다. 데이터의 양이 방대하다는 뜻은 데이터의 사례(instance)가 많은 경우도 있지만, 데이터의

속성의 수가 많은 경우도 있다. 지도학습과 같은 데이터마이닝의 대표적인 알고리즘을 사용하여 방대한 데이터의 패턴을 찾아내고자 할 때, 데이터 속성 중에서 가장 중요한 속성만 선택하여 학습할 수 있다면, 데이터 축소가 이루어져 효율적인 데이터마이닝이 이루어질 수 있다. 그러나 축소된 데이터를 사용할 때 정보의 손실이 나타나는 가가 문제이다. 본 연구를 통해 우리는 적절한 방법에 의한 속성선택을 수행한다면 전체 데이터를 사용하는 것에 비해 오히려 더 좋은 결과를 가져오는 것을 확인할 수 있었다. 특히 연속형 데이터를 이산화 시킨 후 속성선택을 하더라도 원래 데이터를 사용하는 것에 비해 지도학습의 분류 결과가 정확한 것을 알 수 있었고, 속성선택방법 중 MDI, 정보획득비방법이 다른 방법들 보다 약간 더 좋은 결과를 나타냄을 알 수 있었다. 데이터마이닝의 목적이 지도학습을 통해 데이터에 숨겨져 있는 패턴을 발견하고자 하는 것이므로, 연속형 속성을 이산화 시키고 중요한 속성만 골라 데이터를 보관하는 것이 오히려 더 효율적이라는 것을 알 수 있다. 또한 데이터에 잡음이 포함되어 있더라도 이러한 원칙에는 변화가 없는 것을 알 수 있었다. 이 논문에서는 조건속성이 단지 연속형인 경우만 다루었다. 조건속성이 연속형과 이산형으로 혼합되어 있는 경우도 유사한 결과를 가져올 것이라 판단되지만, 이 경우는 더 실험을 하여 입증할 필요가 있다고 생각된다.

참고문헌

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

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

[3] Devijver, P. A. and Kittler J. (1982). *Pattern Recognition: A Statistical Approach*, Prentice Hall International

[4] 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

[5] Hall. M. A. and Holmes, G.(2000). Benchmarking Attribute Selection Techniques for Data Mining(http://www.cs.waikato.ac.nz/~ml)

[6] 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)

[7] Kira, K. and Rendell, L. A.(1992). The feature selection problem : Traditional methods and a new algorithm, *Proceed. of Nat'l Conf. of AI*, 129-134

[8] Kononenko, I.(1994). Estimating attributes : Analysis and extension of RELIEF, *Proceed. of European Conference on Machine Learning*, 171-182

[9] Lee, S. C. and Huh, M. Y.(2003). A Measure of Association for Complex Data, Vol. 44, Issue 1-2, *Computational Statistics and Data Analysis*, 209-220

[10] Liu. H. and Setino, R.(1996). A Probabilistic Approach to Feature Selection: A Filter Solution, In *Proceedings of the 13th International Conference on Machine Learning*, Morgan Kaufmann, 319-327

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

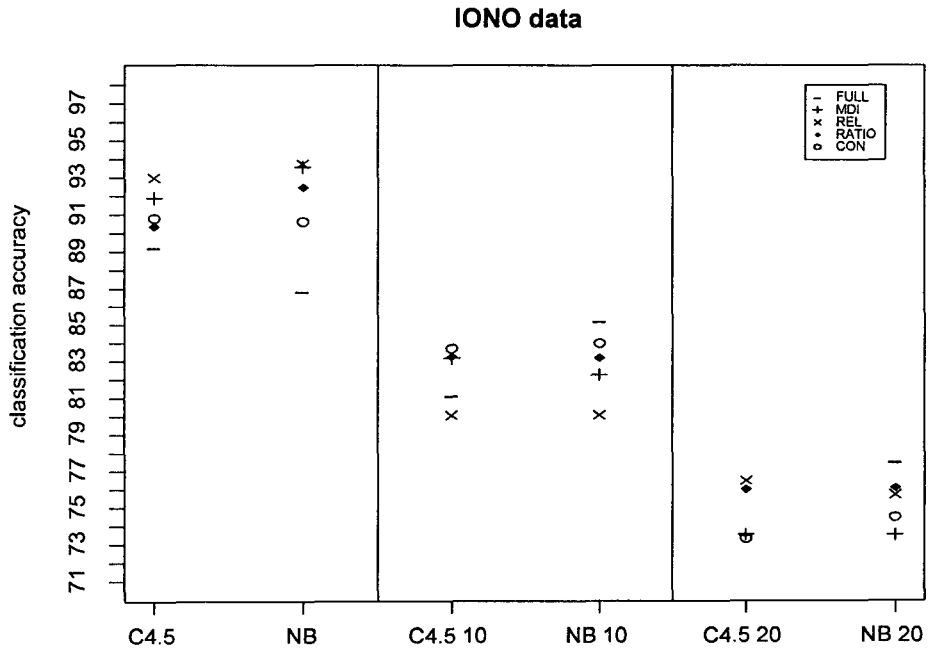
[12] Merz, C. J. and Murphuy, P. M.(1996). UCI Repository of Machine Learning Databases, Department of Information and Computer Science, University of California, Irvine, CA (http://www.ics.uci.edu/~mlearn/MLRepository.html)

[13] Quinlan, J. R.(1986). Induction of decision trees, *Machine Learning*, v.1, 81-106

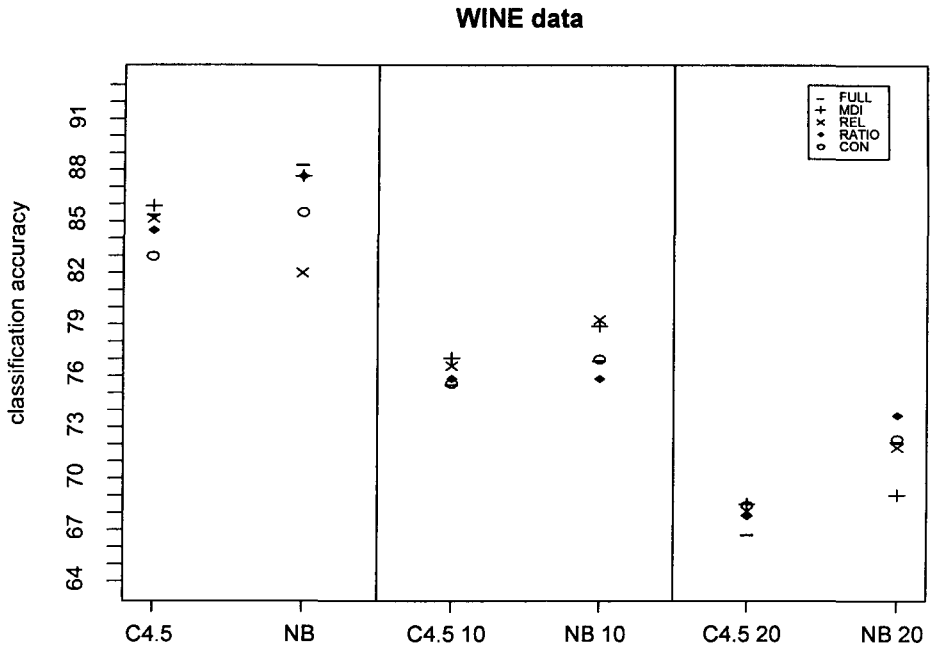
[14] Quinlan. J. R.(1998). *C4.5 : Programs for machine learning*, Morgan Kaufmann Publishers, San Mateo, California

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

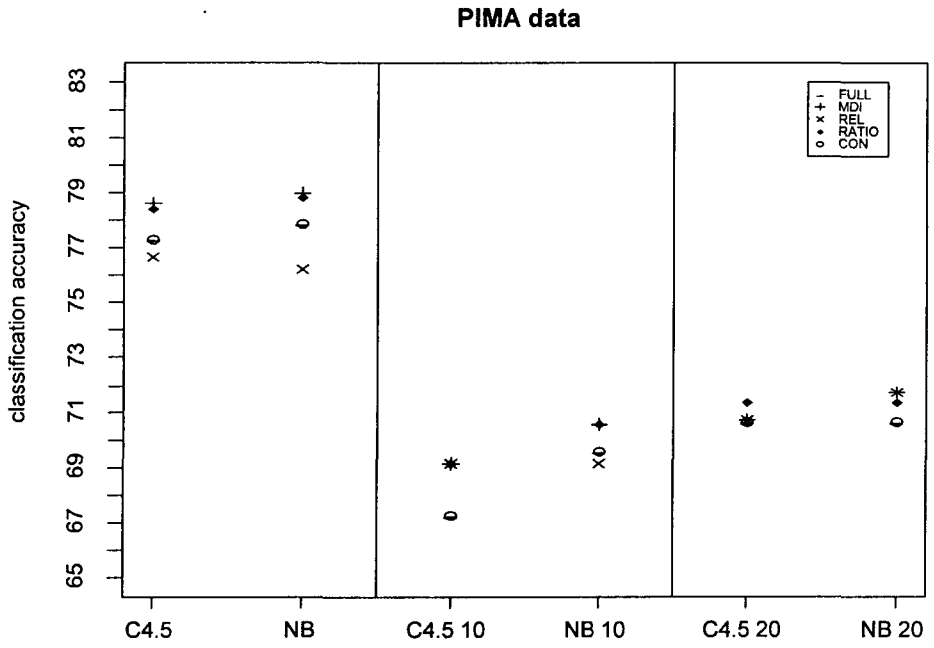
[2003년 9월 접수, 2003년 10월 채택]



[그림1] IONO 데이터의 속성선택방법에 따른 분류정확도



[그림2] WINE 데이터의 속성선택방법에 따른 분류정확도



[그림 3] PIMA 데이터의 속성선택방법에 따른 분류정확도