Journal of the Korean Institute of Industrial Engineers
Vol. 40, No. 4, pp. 375-381, August 2014.
ISSN 1225-0988 | EISSN 2234-6457
http://dx.doi.org/10.7232/JKIIE.2014.40.4.375

<Original Research Paper>

의사결정나무 모델에서의 중요 룰 선택기법

손지은 · 김성범†

고려대학교 산업경영공학과

Rule Selection Method in Decision Tree Models

Jieun Son · Seoung Bum Kim

School of Industrial Management Engineering, Korea University

Data mining is a process of discovering useful patterns or information from large amount of data. Decision tree is one of the data mining algorithms that can be used for both classification and prediction and has been widely used for various applications because of its flexibility and interpretability. Decision trees for classification generally generate a number of rules that belong to one of the predefined category and some rules may belong to the same category. In this case, it is necessary to determine the significance of each rule so as to provide the priority of the rule with users. The purpose of this paper is to propose a rule selection method in classification tree models that accommodate the umber of observation, accuracy, and effectiveness in each rule. Our experiments demonstrate that the proposed method produce better performance compared to other existing rule selection methods.

Keywords: Classification, Decision tree, Data mining, Rule selection

1. 서 론

정보산업의 발전과 더불어 의학, 경제, 경영, 공공 등 다양한 분야에서 데이터가 폭발적으로 쏟아져 나오고 있다(Guo *et al*., 2009). 데이터마이닝 기법은 데이터를 통해 분류, 예측, 군집화, 연관분석 등을 수행하여 가치 있고 흥미로운 정보를 찾아내는 기술이다(Shumeli *et al*., 2010). 그 중, 의사결정나무 모델은 순환적 분할방식을 이용하여 나무모형을 구축한 뒤, 데이터를 분류하거나 예측하는 데이터마이닝 알고리즘으로써 우수한 예측 정확도, 그룹의 세분화, 변수의 중요도 파악, 범주형 변수와 연속 형 변수의 활용 등 수많은 장점을 가지고 있어 다양한 연구 분야에서 널리 사용되고 있다(Mitchell, 1997; Bose *et al*., 2001). 무엇보다도 모델의 결과가 if-then 형식의 룰로 제공되기 때문에, 타 알고리즘에 비해 높은 설명력을 갖는다. 이는 종속변수와 설명변수 사이의 관계를 논리적으로 해석하고자 하는 목적을 갖는 분석에서 유용하게 사용될 수 있다.

위에서 언급한 룰은 종속변수의 예측력을 가장 극대화 시키는 방향으로 결정되며 한 모델에서 다수의 룰이 생성되는 것이 보통이다. 특히, 분류 의사결정나무 모델에서는 같은 범주를 예측하는 룰이 다수 생성될 수 있으며 따라서 각 룰에 대한 중요도를 결정하는 문제가 매우 중요하다. 룰의 중요도란 룰과 해당 룰이 포함하는 관측치 간의 상관정도를 의미하며, 분류 목적에 알맞은 다양한 기준에 따라 평가될 수 있다. 이러한 필요성에도 불구하고 우리가 조사해 본 바로는 의사결정나무 모델에서 생성된 룰의 중요도를 결정하는 연구는 행해지지 않은 것으로 파악되었다.

다만 룰을 생성하는 다른 데이터마이닝 기법에서 룰의 중요도를 결정하는 기법들이 제안되었는데 대표적으로 연관성 분석이다. 연관성 분석에서 사용된 룰 중요도 기법을 의사결정나무 모델에도 그대로 적용해 볼 수 있으나 몇몇 문제점이 발견되었고 따라서 본 연구에서는 이러한 문제점을 보완하여 의사결정나무 모델에 효과적으로 적용할 수 있는 룰 선택기법을 제안하였다.

본 연구는 미래창조과학부의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2013007724). 본 연구는 지식경제부 정보통신 기반 구축사업의 지원을 받아 수행된 연구임(NIPA-2011-(B1110-1101-0002)).
† 연락저자 : 김성범 교수, 136-701 서울시 성북구 안암로 145 고려대학교 산업경영공학과, Tel : 02-3290-3397, Fax : 02-929-5888,
E-mail : sbkim1@korea.ac.kr
2013년 12월 4일 접수; 2014년 4월 2일 수정본 접수; 2014년 4월 29일 게재 확정.

아래 간단한 예제를 통해 기존 룰 중요도 결정 방법의 문제점을 설명해 보도록 하겠다. 기존에는 룰에 대한 중요도를 평가하는 기준으로 분류정확도(Accuracy)를 사용하였다(Safavian *et al*., 1991). 식 (1)에서 N_{all} 은 룰에 해당하는 전체 관측치 개수를 나타내며 N_{corr} 는 룰에 해당하는 관측치 중 옳게 분류된 관측치의 개수를 나타낸다.

$$\text{분류정확도} = N_{corr} \div N_{all} \tag{1}$$

위 식 (1)은 분류정확도는 옳게 분류된 비율만을 고려할 뿐, 옳게 분류된 관측치의 개수 자체는 고려하지 않는 한계점을 가지고 있다. 예를 들어 <Table 1>에서 보여주듯 3개의 룰에 대한 중요도를 결정할 때, 룰 1은 1,000개의 관측치 중에 990개가 올바른 그룹으로 구분되었으나 분류정확도가 룰 2 혹은 룰 3보다 낮으므로 중요도가 낮다고 평가된다. 하지만 룰 1이 분류정확도는 낮더라도 전체 1,011개의 관측치 중에서 약 98%를 차지하는 990개의 관측치를 옳게 분류하고 있기 때문에 룰 1이 다른 룰보다 중요하다고 선택되는 것이 상식적으로 올바른 결정이라고 하겠다.

Table 1. Example of classification accuracy

	N_{all}	N_{corr}	N_{not}	분류정확도
룰 1	1000	990	10	99(%)
룰 2	10	10	0	100(%)
룰 3	1	1	0	100(%)

본 연구에서는 기존 연관규칙 분석에서 사용되는 지지도(Support), 신뢰도(Confidence), 향상도(Lift)의 개념을 바탕으로 룰 중요도 평가 척도를 제안했으며, 실제 데이터를 통해 제안 룰 중요도 평가방법에 대한 유용성을 입증하였다. 즉, 분류정확도에만 의존하여 룰의 중요도를 평가했던 기존 방법과 달리 다양한 정보를 균형 있게 반영하여 룰의 중요도를 평가할 수 있는 척도를 제안하였다. 이는 의사결정나무 모델을 통해 생성된 룰의 중요도를 평가할 수 있는 최초의 기법이라고 하겠다.

이후 본 논문의 구성은 다음과 같다. 제 2장에서는 룰의 중요도 평가와 관련된 기존 연구에 대해 기술하였고, 제 3장은 본 연구에서 제안하는 알고리즘에 대하여 설명하였다. 제 4장은 제안 알고리즘을 실제 데이터에 적용한 결과를 보여주고 기존 방법과 비교하여 성능의 우수성을 입증하였다. 마지막으로 제 5장에서는 본 연구의 요약을 통해 결론을 맺었다.

2. 관련 연구

2.1 Classification Association Rule Mining

Classification Rule Mining(CRM)은 분류 알고리즘을 통해 데이터의 그룹을 분류하는 룰을 찾아내는 기법이다(Quinlan, 1993; Breiman *et al*., 1984). 최근에는 연관성 분석을 통해 생성된 룰을 활용하여 데이터를 분류하는 Classification Association Rule Mining(CARM)에 관한 연구가 활발히 이루어지고 있다. CARM의 대표적인 알고리즘으로는 CBA, CAEP, ADT, CMAR, CPAR 등이 있다(Liu *et al*., 1998; Dong *et al*., 1999; Wang *et al*., 2000; Li *et al*., 2001; Han, 2003). CARM에서는 그룹을 분류하는 중요한 룰을 찾기 위해 다음과 같은 두 단계의 과정을 거친다. 먼저 1단계에서는 데이터로부터 연관규칙을 생성하며 2단계에서 생성된 룰을 중요도 평가기준에 따라 정렬한다.

2.2 룰 정렬 방식

(1) '지지도-신뢰도' 프레임워크

단계 1을 통해 생성된 룰을 'Rule : X → Y'이라고 표현할 때 X를 조건부, Y를 결과부라고 지칭한다. 룰에 대한 해석은 'X 조건에 만족하는 관측치는 Y 그룹으로 분류된다'이며 룰에 대한 중요도 평가 척도로 지지도, 신뢰도, 그리고 향상도가 있다. 이는 연관분석 알고리즘 중 가장 대표적인 Apriori 알고리즘에 적용되고 있으며 가장 일반적으로 쓰이는 룰 중요도 평가 척도이다(Agrawal *et al*., 1993; Agrawal *et al*., 1994).

$$\text{지지도} = \Pr(X \cup Y) \tag{2}$$

$$\begin{aligned}\text{신뢰도} &= \Pr(X \cup Y) \div \Pr(X) \\ &= \Pr(Y \mid X)\end{aligned} \tag{3}$$

$$\begin{aligned}\text{향상도} &= \Pr(X \cap Y) \div (\Pr(X) \times \Pr(Y)) \\ &= \text{신뢰도} \div \Pr(Y)\end{aligned} \tag{4}$$

식 (2)의 지지도는 전체 관측치 중에서 조건부와 결과부를 모두 만족하는 관측치의 비율을 의미한다. 식 (3)의 신뢰도는 조건부를 만족하는 관측치 중에서 조건부와 결과부를 동시에 만족하는 관측치의 비율이다. 즉, 조건부 확률을 통해 조건부와 결과부의 상관관계에 대한 평가에 초점을 두고 있다. 식 (4)의 향상도는 지지도와 신뢰도만으로 룰의 중요도를 판단하기 어려울 때 고려되는 평가 척도로써 조건부가 주어지지 않고 결과부를 만족할 확률 대비, 조건부가 주어졌을 때 결과부를 만족할 확률이다. 향상도의 값이 1이라는 것은 조건부와 결과부가 상호 독립적인 관계이고 1보다 크면 양의 상관관계, 1보다 작으면 음의 상관관계를 갖는다.

'지지도-신뢰도' 프레임워크는 지지도, 신뢰도, 조건부의 크기를 동시에 평가 척도로 이용하여 룰을 정렬하는 방식으로 크게 두 가지 방법이 있다(Wang *et al*., 2007). 먼저, Confidence-Suport-size_of_Antecedent(CSA)는 생성된 룰을 먼저 신뢰도가 높은 순서대로 내림차순 정렬을 한 후 동일한 신뢰도 값을 갖은 룰에 대해서 지지도가 높은 순서대로 내림차순 정렬을 한다. 최종적으로 동일한 신뢰도와 지지도를 갖는 룰에 대해서 조건부의 크기가 작은 것부터 오름차순 정렬한다. 두 번째 방법인 size_of_Antecedent-Confidence-Support(ACS)는 조건부 크기를

먼저 고려하여 정렬한 후 동일한 조건부 크기를 갖는 룰에 대해서 신뢰도와 지지도를 차례대로 내림차순 정렬한다. <Table 2>를 통해 CSA와 ACS 방식간의 중요도 평가 결과의 차이를 확인할 수 있다.

Table 2. Rule selection results (a) CSA method and (b) ACS method

(a)

중요도순위	신뢰도	지지도	조건부 크기	룰 번호
1	100	100	5	룰 1
2	100	90	5	룰 2
3	90	80	2	룰 3
4	90	80	7	룰 4
5	70	90	5	룰 5

(b)

중요도순위	조건부 크기	신뢰도	지지도	룰 번호
1	2	90	80	룰 3
2	5	100	100	룰 1
3	5	100	90	룰 2
4	5	70	90	룰 5
5	7	90	80	룰 4

(2) 가중 상대적 분류정확도

상대적 분류정확도(Relative Accuracy)는 해당 룰 내에서만 분류정확도를 계산하는 기존의 분류정확도의 문제점을 보안하기 위해 고안된 기법으로 식 (5)와 같다(Lavrač *et al*., 1999).

$$\text{상대적 분류정확도} = \Pr(Y|X) - \Pr(Y) \tag{5}$$

$\Pr(Y|X)$는 기존의 분류정확도를 나타내며 $\Pr(Y)$은 전체 데이터에서 결과부를 만족하는 관측치의 비율이다. 즉, 전체에서 결과부를 만족하는 관측치의 비율이 크면 상대적으로 룰의 분류정확도를 낮게 계산하는 방법이다. 그러나 그룹 별 관측치의 개수가 균등하게 분포된 데이터일 경우 기존 분류정확도와의 차이가 없기 때문에 이러한 문제점까지 보완하여 고안된 기법이 가중 상대적 분류정확도(Weighted Relative Accuracy : WRACC)이며 식 (6)과 같다(Lavrač *et al*., 1999).

$$\begin{aligned}\text{가중 상대적 분류정확도} &= \Pr(X) \times (\Pr(Y|X) - \Pr(Y)) \\ &= \Pr(Y|X) \times \Pr(X) \\ &\quad - \Pr(Y) \times \Pr(X)\end{aligned} \tag{6}$$

여기서, 가중치 $\Pr(X)$는 룰에 의해 설명될 수 있는 관측치의 비율로써 그 값이 클수록 가중치가 높아져 룰의 중요도가 높아진다.

(3) 라플라스

앞에서 언급한 방법들에서 사용하고 있는 분류정확도는 옳게 분류된 비율만을 고려하였을 뿐, 옳게 분류된 관측치의 개수 자체는 전혀 고려하지 않는다. 예를 들어, <Table 1>에서 룰 2와 룰 3은 각기 옳게 분류하는 관측치의 개수가 다르지만 분류정확도는 같다. 이와 같은 한계를 극복하고자 CARM 에서는 라플라스 방법을 적용하여 옳게 분류된 비율뿐만 아니라 관측치 개수까지 고려하였다(Lavrač *et al*., 1999; Coenen *et al*., 2004; Wang *et al*., 2007). 식 (7)에서 K는 데이터의 클래스 개수이다.

$$\text{라플라스} = (N_{corr} + 1) \div (N_{all} + K) \tag{7}$$

라플라스는 의사결정나무 생성 시, 가지치기 알고리즘으로 가장 널리 쓰이는 엔트로피 알고리즘의 문제를 보완하기 위하여 처음 고안되었다(Clark *et al*., 1989). 즉, 분류정확도가 높을지라도 룰이 포함하는 전체 관측치 개수가 적으면 라플라스 값은 낮아지게 된다. <Table 3>은 세 개의 룰 모두 분류정확도는 동일하게 100%를 나타내고 있다. 그러나 라플라스 적용 결과, 포함하는 관측치의 개수가 적은 룰 3의 경우 66.6%로 떨어진다.

Table 3. Performance of Laplace method(k = 2) for a toy example

	N_{all}	N_{corr}	N_{not}	분류 정확도	라플라스
룰 1	100	100	0	100%	99%
룰 2	10	10	0	100%	91.6%
룰 3	1	1	0	100%	66.6%

이처럼 관측치 개수와 분류정확도를 동시에 고려하여 룰을 평가할 수 있다는 장점 때문에 다양한 CRAM 관련 연구에서 활용되어져 왔다. 그러나 애초에 라플라스를 고안한 목적은 아주 적은 개수의 관측치를 갖는 룰에 대해서만 중요도를 낮추는 것이다. 즉, 일정 수준의 관측치 개수가 확보된 룰에 대해서는 기존의 분류정확도와의 차이가 거의 없어진다.

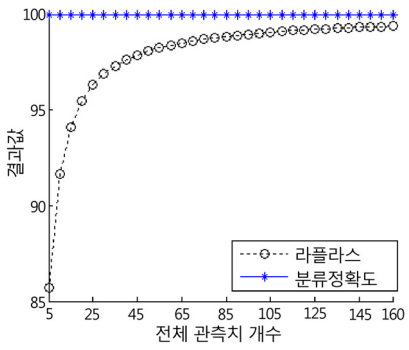


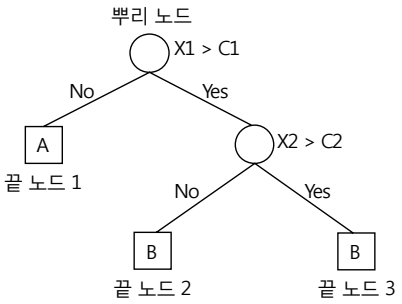
Figure 1. Performance (classification accuracy) of Laplace method over different numbers of observations

<Figure 1>은 기존의 분류정확도가 100%일 경우, 라플라스와 비교해 보았다. X축은 룰에 해당하는 전체 관측치의 개수

이며 Y축은 기존 분류정확도 및 라플라스의 값이다. 그래프를 통해 룰이 포함하고 있는 전체 관측치 개수에 따른 라플라스 값의 변화를 살펴볼 수 있다. 관측치 개수가 25개 미만일 경우에는 기존 분류정확도와 라플라스의 차이가 확연히 들어나는 것을 확인할 수 있으며, 이는 모든 관측치를 옳게 구분하여 분류정확도가 100%일지라도 전체 관측치의 개수가 약 25개 이하일 때에는 낮은 라플라스 결과 값을 갖게 된다. 그러나 25개 이상의 관측치를 포함하는 룰에 대해서는 모두 95% 이상의 높은 라플라스 결과 값을 갖게 되며 관측치의 개수가 증가하면 할수록 기존 분류정확도와 라플라스 결과 값의 차이가 거의 나타나지 않아 라플라스 적용의 의미가 사라지게 된다.

3. 제안 알고리즘

<Figure 2>(a) 는 예시를 통해 의사결정나무 모델의 간단한 형태를 보여주고 있다.



(a)

룰 1	'X1 < = C1' → A category
룰 2	'X1 > C1' & 'X2 < = C2' → B category
룰 3	'X1 > C1' & 'X2 > C2' → B category

(b)

Figure 2. Overview of a decision tree model and a list of rules generated from the model

의사결정나무 모델을 통해 생성된 룰이란 뿌리 노드로부터 각 끝노드까지의 경로를 if-then의 형식으로 나타낸 것이다. <Figure 2(b)>는 <Figure 2(a)>를 통해 생성된 3개의 룰의 if-then의 형식이다. 룰의 조건부는 끝노드를 설명하는 변수와 기준 값으로 이루어져 있으며 결과부는 해당 끝노드가 분류된 그룹의 범주로 이루어져 있다. 의사결정나무 알고리즘을 통해 생성된 각 룰에 대해서 가중치, 지지도, 신뢰도, 향상도 값을 생성한다. <Table 4(a)>는 <Figure 2>의 룰에 대한 그룹 범주와 관측치 개수를 나타내고 있는 예이다. 가중치는 전체 관측치 중에서 조건부가 포함하는 관측치의 비율이며 <Table 4(b)>와 같은 결과를 나타낸다. 룰 1의 경우 전체 1,000개의 관측치 중에서 조건부에 해당하는 관측치는 500이므로 0.5의 가중치를 갖게 된다.

Table 4. (a) The number of observations that correctly classified in each rule, (b) Calculation of weight in each rule

(a)

	범주	N_{all}	N_{corr}	N_{not}
룰 1	A	500	400	100
룰 2	B	300	200	100
룰 3	B	200	100	100

(b)

	전체 관측치 개수	N_{all}	가중치
룰 1	1,000	500	0.5
룰 2	1,000	300	0.3
룰 3	1,000	200	0.2

지지도는 전체 관측치 중에서 조건부와 결과부를 동시에 포함하는 관측치의 비율이며 <Table 5(a)>와 같은 결과를 나타낸다. 룰 1의 경우 조건부와 결과부를 동시에 포함하는 데이터의 개수는 'X1 < = C1'조건을 만족하는 데이터 중, A 클래스로 구분된 관측치의 개수이다. 신뢰도는 조건부를 포함하는 관측치 중에서 조건부와 결과부를 동시에 포함하는 관측치의 비율이며 <Table 5(b)>와 같은 결과를 나타낸다. 즉, 룰이 포함하는 관측치 중에서 옳게 분류한 관측치의 비율을 뜻하는 것으로 기존의 분류정확도와 그 의미가 같다. 향상도는 조건부가 없이 결과부를 선택할 확률 대비 조건부가 주어졌을 때 결과부를 선택할 확률이며 <Table 5(c)>와 같은 결과를 나타낸다. 룰 1의 경우 전체 관측치 중에서 조건부 없이 A 클래스를 선택할 확률은 600 ÷ 1000 = 0.6이다. 그리고 조건부가 주어졌을 때 결과부를 선택할 확률은 400 ÷ 500 = 0.8로 앞서 산출된 신뢰도값과 동일하다. 예를 통해 생성된 3개 룰의 향상도는 모두 1 이상이므로 연관성이 높은 룰이라 할 수 있다.

Table 5. Calculation of (a) support, (b) confidence, and (c) lift

(a)

	전체 관측치 개수	N_{corr}	지지도
룰 1	1,000	400	40%
룰 2	1,000	200	20%
룰 3	1,000	100	10%

(b)

	N_{all}	N_{corr}	신뢰도
룰 1	500	400	80%
룰 2	300	200	67%
룰 3	200	100	50%

(c)

	Pr(그룹)	신뢰도	향상도
룰 1	600/1,000	80%	1.33
룰 2	400/1,000	67%	1.67
룰 3	400/1,000	50%	1.25

마지막으로 제안방법의 척도인 Weighted Support-Confidence-Lift(WSCL)은 각 룰로부터 계산된 가중치, 지지도, 신뢰도, 향상도를 이용하여 식 (8)과 같이 계산한다.

$$WSCL = \text{가중치} \times \text{지지도} \times \text{신뢰도} \times \text{향상도} \quad (8)$$

식 (8)을 통해 산출된 WSCL은 <Table 6>와 같다.

Table 6. Calculation of WSCL

	가중치	지지도	신뢰도	향상도	WSCL
룰 1	0.5	40%	80%	1.33	0.212
룰 2	0.3	20%	67%	1.67	0.067
룰 3	0.2	10%	50%	1.25	0.012

4. 실험 및 결과

4.1 데이터

제안하는 알고리즘의 성능을 입증하기 위해 실제 데이터에 적용해 보았다. 실험에서 사용된 데이터 셋은 총 7개이며, 모두 UCI ML Repository(http://archive.ics.uci.edu/ml/)로부터 얻은 데이터 셋이다. 그 중, 'Adult' 데이터 셋은 상세한 실험 결과를 설명하기 위해 사용되었다. 'Adult' 데이터 셋은 나이, 학력, 직업, 인종, 성별, 국적 등의 13개 설명변수를 통해 연간수입이 5만 달러를 넘는지 아닌지를 구분하는 것을 목적으로 하고 있으며 총 관측치 개수는 3,571개이다.

4.2 결과

의사결정나무 알고리즘 중 CART 알고리즘을 사용하여 의사결정나무를 생성하였다(Breiman *et al*., 1984). 전체 데이터 셋을 70 : 30으로 나누어 각각 학습데이터 셋과 테스트 데이터 셋으로 활용하였으며 Depth는 학습 에러와 테스트 에러를 최소화 시키는 '5'로 설정하였다. 의사결정나무 생성 결과 학습 에러는 84.4%, 테스트 에러는 85.3%를 나타내었으며, <Table 7>과 같이 총 8개의 끝노드를 생성하였다.

Table 7. Generated rules by the decision tree model with 'Adult' dataset

	범주	N_{all}	N_{corr}	N_{not}
룰 1	B	1,486	1,400	86
룰 2	A	80	80	0
룰 3	B	641	464	177
룰 4	A	40	36	4
룰 5	A	47	47	0
룰 6	A	256	166	90
룰 7	B	105	73	32
룰 8	A	92	53	39

각 룰별로 가중치, 지지도, 신뢰도, 향상도를 산출하고 이 값들을 모두 곱하여 최종적으로 WSCL를 산출하였다. <Table 8>은 각 룰에 대한 가중치, 지지도, 신뢰도, 향상도이며 <Table 9>는 WSCL와 그에 따른 중요도 순위를 나타내고 있다.

Table 8. Weight, support, confidence, and lift

	가중치	지지도	신뢰도	향상도
룰 1	0.54	50.96	94.21	1.25
룰 2	0.03	2.91	100	4.12
룰 3	0.23	16.89	72.39	0.96
룰 4	0.01	1.31	90	3.71
룰 5	0.02	1.71	100	4.12
룰 6	0.09	6.04	64.84	2.67
룰 7	0.04	2.66	69.52	0.92
룰 8	0.03	1.93	57.61	2.37

Table 9. Importance of rules based on WSCL

	WSCL	중요도 순위
룰 1	3,247.40	1
룰 2	34.93	4
룰 3	274.07	2
룰 4	6.37	8
룰 5	12.06	5
룰 6	97.52	3
룰 7	6.52	7
룰 8	8.83	6

기존의 룰 중요도 평가척도인 라플라스와 WRACC의 실험결과와 본 연구에서 제안하는 WSCL의 실험결과를 비교하였다. <Table 10>은 각 평가척도를 통한 중요도 순위이다.

Table 10. Comparison of rule selection results among Laplace, WRACC, and WSCL(proposed)

	라플라스	WRACC	WSCL (Proposed)
룰 1	3	1	1
룰 2	1	3	4
룰 3	5	8	2
룰 4	4	6	8
룰 5	2	4	5
룰 6	7	2	3
룰 7	6	7	7
룰 8	8	5	6

각 평가 척도로부터 결정된 룰의 중요도 순위가 실제로 분류정확도, 포함하는 관측치 개수, 룰의 유효성 등을 균형 있게 반영하였는지에 대해 판단하기 위해 <Figure 3>과 같은 그래프를 그려보았다. X축은 각 기법을 통해 선택된 룰을 중요도

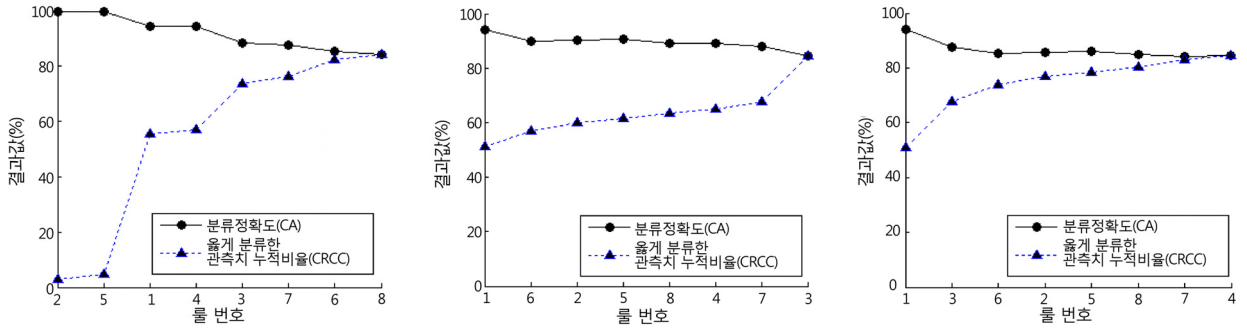


Figure 3. The performance comparison of three methods representing the classification accuracy(CA) and cumulative ratio of correctly classified observations(CRCC) with the different number of rule sets. : The result of (a)Laplace, (b)WRACC, and (c)WSCL (proposed)

에 따라 정렬하였으며 Y축은 각 룰에 대한 데이터의 분류정확도와 옳게 분류한 관측치의 누적비율이다. <Figure 3(a)>에서 보면 라플라스 평가 척도를 통해 중요하다고 선택된 룰은 분류정확도는 높지만 옳게 분류되어지는 관측치의 비율은 매우 낮다. 하지만 해당 룰에 속한 관측치의 비율을 고려한 WRACC의 경우, <Figure 3(b)>와 같이 비교적 분류정확도와 관측치 비율을 균형 있게 고려하여 룰의 중요도를 선정한 것을 볼 수 있다. 그러나 WRACC는 잘못 분류된 관측치의 개수까지 포함하여 가중치를 주기 때문에 단순히 관측치를 많이 포함하고 있는 룰이 중요하다고 선택 될 수 있다. <Figure 3(c)>는 본 연구에서 제안하는 WSCL을 통한 결과로써 분류정확도와 옳게 분류한 관측치 비율이 위의 두 방법보다 균형 있게 고려되었다는 것을 확인할 수 있다. 즉, 분류정확도와 옳게 분류한 관측치의 누적비율을 적절히 고려하여 산출된 결과는 두 값의 차이가 작아야 한다. 따라서 각 그래프에서 두 곡선에서 점과 점 사이의 거리를 계산하여 총 합을 구한 결과, 라플라스는 298, WRACC는 206, WSCL은 96으로 제안하는 알고리즘인 WSCL에서 거리의 합이 가장 작았다. 제안하는 평가 척도인 WSCL 성능의 견고함을 평가하기 위해, 보다 많은 데이터에 위와 동일한 과정의 실험을 적용하였다. <Table 11>은 앞서 설명한 'Adult' 데이터 셋을 포함한 총 7개의 데이터 셋에 대하여 각 평가척도별 거리를 나타내고 있다. 정확한 비교를 하기위해서, 거리는 각 데이터의 룰의 개수에 따라 정규화 시킨 결과를 나타내었다. 7개 중, 4개의 데이터 셋에서 WSCL의 거리가 가장 짧으며 이는 WSCL을 통한 중요 룰 선택 결과가 예측 정확도와 옳게 분류한 관측치 개수를 가장 균형 있게 반영한 결과라 해석할 수 있다. 7개 중, 3개의 데이터 셋에서는 WRACC의 거리가 가장 짧으나 WSCL의 결과와 거의 차이가 없다. 7개 데이터 셋에 대한 각 평가 척도 별 평균 거리 비교 결과 WSCL, 라플라스, WRACC 순서로 거리가 짧았다. 즉 WSCL의 거리가 기존의 평가척도와 비교했을 때, 평균적으로 거리가 짧으므로 이는 제안 알고리즘의 효용성과 우수성을 보여주는 결과라고 하겠다.

Table 11. The comparison of Euclidean distances from CA to CRCC among Laplace, WRACC, and WSCL(proposed)

	라플라스	WRACC	WSCL (Proposed)
Data 1	37.3	25.8	12
Data 2	6.7	8.9	6.6
Data 3	26.2	18.8	16.3
Data 4	28.2	18.8	19.1
Data 5	29.5	70.7	13.5
Data 6	19.8	11.5	13.1
Data 7	18.2	17.5	17.6
평균	23.7	24.6	14.0

5. 결 론

데이터마이닝 모델에서 가장 널리 쓰이고 있는 의사결정나무 모델은 분류 문제에서 높은 분류정확도를 나타낼 뿐만 아니라 if-then 형식의 룰을 제공하여 효과적인 해석이 가능하다. 또한, 뛰어난 해석력이 장점인 만큼 생성된 룰의 정확한 중요도 파악을 반드시 필요하다. 그러나 각 룰이 분류정확도, 포함하고 있는 관측치 개수, 통계적 유의성 등 수많은 정보를 담고 있음에도 불구하고 기존의 연구에서는 분류정확도만을 사용하여 룰의 중요도를 평가하였다. 연관성 분석에서 룰 평가 척도로 사용되는 라플라스와 WRACC를 의사결정나무 모델을 통해 생성된 룰에 적용하여 중요도를 평가하였으나 다양한 정보를 균형 있게 반영하는데 한계점이 존재하였다.

본 연구에서는 관측치의 개수와 예측력, 그리고 유효성 등 룰의 다양한 정보를 균형 있게 반영하여 룰의 중요도를 평가할 수 있는 척도를 제안하였다. 실제 데이터에 적용한 결과, 기존 라플라스와 WRACC 기법보다 의사결정나무로부터 생성된 룰의 특성을 잘 고려한 선택을 했음을 확인할 수 있었다. 본 연구결과는 의사결정나무를 활용하는 연구 및 실무 사례에서

효율적이고 정확한 결과해석 및 의사결정에 도움을 줄 수 있을 것이다. 하지만 본 연구는 분류정확도와 관측치의 개수를 통해서만 제안하는 방법의 성능을 평가하였다. 추후 룰의 유효성 및 예측정확도 등 다양한 관점에서 알고리즘의 일반적인 효용성에 대한 평가가 이루어져야 하겠다.

참고문헌

Agrawal, R., Imieliński, T., and Swami, A. (1993), Mining association rules between sets of items in large databases, *In ACM SIGMOD Record*, **22**(2), 207-216.

Agrawal, R. and Srikant, R. (1994), Fast algorithms for mining association rules, *In Proc. 20th Int. Conf. Very Large Data Bases*, VLDB, 1215, 487-499.

Bose, I. and Mahapatra, R. K. (2001), Business data mining-a machine learning perspective, *Information and management*, **39**(3), 211-225.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), In *Classification and regression trees Belmont*, CA : Wadsworth International Group.

Clark, P. and Niblett, T. (1989), The CN2 induction algorithm, *Machine learning*, **3**(4), 261-283.

Coenen, F. and Leng, P. (2004), An evaluation of approaches to classification rule selection, *In Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on*, 359-362.

Dong, G., Zhang, X., Wong, L., and Li, J. (1999), CAEP : Classification by aggregating emerging patterns, *In Discovery Science*, 30-42.

Guo, Z., Singh, R., and Pierce, M. (2009), Building the Polar Grad Portal Using Web 2.0 and Open Social, *Pervasive Technology Institute Indiana University*, Bloomington, Indiana.

Han, J. (2003), CPAR : Classification based on predictive association rules, *In Proceedings of the third SIAM international conference on data mining*, **3**, 331-335.

Lavrač, N., Flach, P., and Zupan, B. (1999), *Rule evaluation measures : A unifying view*, 174-185.

Li, W., Han, J., and Pei, J. (2001), CMAR: Accurate and efficient classification based on multiple class-association rules, *In Data Mining, 2001, ICDM 2001, Proceedings IEEE International Conference on*, 369-376.

Liu, B., Hsu, W., and Ma, Y. (1998), Integrating classification and association rule mining, *In Proceedings of the 4th*.

Mitchell, T. M. (1997), Machine Learning, 52-78, Singapore, The McGraw-Hill Companies Inc..

Quinlan, J. R. (1993), C4. 5 : Programs for machine learning, Morgan Kaufmann.

Safavian, S. R. and Landgrebe, D. (1991), Asurvey of decision tree classifier methodology, *Systems*, *Man and Cybernetics*, *IEEE Transactions on*, **21**(3), 660-674.

Shumeli, G., Patel, N. R., and Bruce, P. C. (2010), Data Mining for Business Intelligence, 2nd ed, WILEY, Canada, 3-38.

Wang, Y. J., Xin, Q., and Coenen, F. (2007), A novel rule ordering approach in classification association rule mining, *In Machine Learning and Data Mining in Pattern Recognition*, 339-348.

Wang, K., Zhou, S., and He, Y. (2000), Growing decision trees on support-less association rules, *In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 265-269.