Journal of The Korea Society of Computer and Information
Vol. 22 No. 1, pp. 23-32, January 2017
www.ksci.re.kr
https://doi.org/10.9708/jksci.2017.22.01.023

Subgroup Discovery Method with Internal Disjunctive Expression

Seyoung Kim*, Kwang Ryel Ryu**

Abstract

We can obtain useful knowledge from data by using a subgroup discovery algorithm. Subgroup discovery is a rule model learning method that finds data subgroups containing specific information from data and expresses them in a rule form. Subgroups are meaningful as they account for a high percentage of total data and tend to differ significantly from the overall data. Subgroup is expressed with conjunction of only literals previously. So, the scope of the rules that can be derived from the learning process is limited. In this paper, we propose a method to increase expressiveness of rules through internal disjunctive representation of attribute values. Also, we analyze the characteristics of existing subgroup discovery algorithms and propose an improved algorithm that complements their defects and takes advantage of them. Experiments are conducted with the traffic accident data given from Busan metropolitan city. The results shows that performance of the proposed method is better than that of existing methods. Rule set learned by proposed method has interesting and general rules more.

▸ Keyword : Data mining, Subgroup discovery, Rule learning, Traffic accident data

I. Introduction

우리는 데이터 분석을 통해 현상을 관찰하고 유용한 정보를 찾을 수 있다. 획득한 정보는 사회 정책 수립 혹은 기업의 제품 판매 전략 등 다양하게 이용될 수 있다. 데이터를 분석하는 방법에는 통계적 분석 방법과 기계 학습 알고리즘을 이용한 방법이 있다. 기계 학습 알고리즘은 분석 목적에 따라 다양한데 본 논문에서는 데이터로부터 특이한 경향을 발견하고 그것을 규칙 형태로 표현해주는 서브그룹 디스커버리 알고리즘에 관해 다루고자 한다.

서브그룹 디스커버리는 일반적 경향과 현저히 다른 양상을 보이는 데이터 그룹을 찾는 것 이다. 즉, 데이터가 주어졌을 때 전체 데이터에서 분석 목표로 삼은 클래스 데이터의 비율을 목표 클래스의 일반적 경향이라고 보고 목표 클래스 데이터의 비율이 현저히 높은 데이터 그룹을 찾는 것이다. 이렇게 찾은 데이터 그룹은 해당 데이터들의 공통적인 특성을 조건으로 하여 'if 특성 then 목표 클래스'와 같은 규칙의 형태로 표현을 할 수 있다. 교통사고 데이터를 이용하여 음주운전 사고에 대해 분석을 한다고 가정해보자. 전체 교통사고 데이터 중 음주운전 사고 데이터가 1% 를 차지하는데 주말에 발생한 교통사고 데이터 중 음주운전 사고데이터의 비율이 3% 이라면 주말에 음주운전 사고 발생 확률이 높다고 할 수 있다. 따라서 주말에 사고가 발생한 데이터 그룹을 'if 주말 then 음주운전 사고'이라고 표현할 수 있다.

그러나 데이터 그룹의 크기가 너무 작아서 비율이 현저히 높아진 경우 데이터의 특성을 설명한다고 볼 수 없다. 교통사고 데이터 중 주말에 발생한 사고가 1건(즉, 데이터 그룹의 크기가 1)이고 그 하나의 데이터가 음주운전 사고라고 가정하자. 그러면 주말에 발생한 교통사고 데이터 중 음주운전 사고의 비율이 100%가 되지만 1건의 데이터를 바탕으로 주말과 음주운전 사

• First Author: Seyoung Kim, Corresponding Author: Kwang Ryel Ryu
*Seyoung Kim(birdzero@pusan.ac.kr), Dept. of Computer Science and Engineering, Pusan National University
**Kwang Ryel Ryu(krryu@pusan.ac.kr), Dept. of Computer Science and Engineering, Pusan National University
• Received: 2016. 10. 11, Revised: 2016. 11. 24, Accepted: 2016. 12. 27.
• This work was supported by BK21PLUS, Creative Human Resource Development Program for IT Convergence.

고가 깊은 관계가 있다고 말하기는 어렵다. 따라서 서브그룹 디스커버리는 규칙의 특이성뿐만 아니라 일반성 또한 중요하다.

본 논문에서는 규칙의 조건부에 속성 값의 선언적 표현이 가능한 서브그룹 디스커버리 방안을 제안한다. 조건부에 속성 값을 선언적으로 표현할 수 있으면 표현력의 증가로 인해 커버할 수 있는 데이터 그룹의 종류가 다양해진다. 즉, 기존의 방식보다 다양한 규칙을 찾을 수 있다. 또한 규칙이 커버하는 데이터 그룹의 크기 또한 증가할 수 있으므로 규칙의 일반성 또한 확대할 수 있다.

대표적인 서브그룹 디스커버리 알고리즘으로 MIDOS와 CN2-SD가 있다. 이 알고리즘들은 규칙의 일반성과 특이성을 함께 고려하기 위해 가중 상대 정확도(weighted relative accuracy, WRA)를 측정하여 규칙을 학습한다 [1].WRA에서 가중치(weight)는 전체 데이터 중 규칙의 조건부를 만족하는 데이터의 비율을 의미하며 결국 규칙의 일반성을 나타낸다. 상대정확도(Relative accuracy)는 규칙에 대응하는 데이터 그룹에서 목표 클래스가 차지하는 비율, 즉 규칙의 정확도와 전체 데이터에서 차지하는 목표 클래스의 비율의 차로 정의되며 규칙의 특이성을 나타낸다. WRA는 이 두 가지의 곱으로 측정할 수 있다.

MIDOS [1]는 WRA를 기준으로 최상위 k개의 규칙을 학습하는 알고리즘이다. 사용자의 정의에 따라 k개의 규칙을 한 번에 학습할 수 있지만 규칙에 의해 커버되는 데이터가 유사하고 규칙의 조건부 또한 유사한 규칙들을 다량 학습한다. k값이 커질수록 다양한 규칙들을 학습할 가능성이 커지지만 규칙의 수가 너무 많아지면 규칙을 분석하기 어려워진다는 문제점이 있다. [2]에서는 이러한 문제점의 해결방안으로 학습을 통해 얻은 규칙들을 세 가지 단계의 후처리를 통해 간소화 하는 방안을 제안한다. 후처리를 통해 규칙의 수를 상당히 줄일 수 있지만 이를 위해 사용자가 정해야 할 계수들이 많다는 단점이 있다. 본 논문에서는 내부 디스정션으로 표현된 규칙을 학습하면서 그로 인해 확장된 규칙 집합의 크기를 줄이기 위한 방안을 제안하며 사용자 정의 계수를 줄이고 후처리 과정을 간결하게 하고자 한다.

한 편, CN2-SD는 목표 클래스에 해당되는 모든 데이터를 커버할 때까지 반복하며 각 라운드를 거칠 때마다 규칙을 하나씩 학습하는 알고리즘이다 [3],[4].모든 데이터를 커버하는 것은 너무 엄격하기 때문에 조건을 완화시킨 변종 알고리즘들이 있다. 모든 데이터를 커버하기 위한 규칙의 집합을 학습하기 때문에 MIDOS와는 달리 다양한 규칙을 학습할 수 있다는 장점이 있지만 커버되지 못한 데이터를 커버하는 것에 집중하여 규칙을 학습하기 때문에 개개의 규칙의 질이 좋지 않을 수 있다. 심지어 반복을 거듭하다 보면 목표 클래스의 일반적 경향에도 미치지 못하는, 다시 말해 목표 클래스의 특성과 관계 없는 규칙이 학습되기도 한다.

본 논문에서는 이러한 두 알고리즘의 장점을 취하고 단점을 보완하여 학습된 규칙의 질을 향상시킬 수 있도록 두 알고리즘을 결합한 형태의 서브그룹 디스커버리 방안을 제안한다. CN2-SD 알고리즘에서 규칙을 하나씩 학습하며 목표 클래스의 데이터의 일정 비율을 커버할 때까지 규칙을 찾았다면 제안하는 알고리즘은 MIDOS 알고리즘을 통해 한 라운드에 여러 개의 규칙을 학습하면서 목표 클래스의 데이터를 일정 비율만큼 커버할 때까지 반복한다. 따라서 MIDOS 알고리즘만으로는 찾을 수 없었던 새로운 측면의 규칙을 학습할 수 있으면서 데이터의 특성을 살펴볼 수 있는 질 좋은 규칙들을 학습 할 수 있다.

본 논문의 구성은 다음과 같다. 다음 장에서는 앞서 언급한 기존 방식의 서브그룹 디스커버리 알고리즘에 대해 자세히 설명한다. 3장과 4장에서는 본 논문에서 제안하는 내부 디스정션으로 표현된 규칙을 획득하는 방안과 서브그룹 디스커버리 알고리즘을 차례로 소개한다. 5장에서는 실험 결과로써 제안한 방안의 성능을 정량적으로 평가한 결과와 더불어 정성적인 평가 결과를 보이고 6장에서 결론을 맺는다.

II. Related Works

서브그룹 디스커버리는 클래스 분별이 가능한 데이터를 이용하여 대상 클래스에 대한 규칙을 학습하는 지도 학습(supervised learning) 기법과 정해진 대상 클래스 없이 속성 간의 연관성을 분석하여 규칙을 학습하는 자율 학습(unsupervised learning) 기법이 있다 [6], [7]. 서브그룹 디스커버리 알고리즘은 분류를 위한 규칙 학습 알고리즘을 확장한 알고리즘 [3],연관 규칙 학습 알고리즘을 확장한 알고리즘 [5], 진화형 알고리즘을 적용하여 규칙을 학습하는 알고리즘 [8]등이 있다. 학습된 규칙 집합은 데이터의 특성을 파악하기 위한 것이므로 서술 모형(descriptive model)이다. 규칙 집합은 분류 혹은 예측을 위한 모델로서도 작동할 수 있지만 이를 위해서는 데이터에 대한 다양한 분석이 아닌 분류 혹은 예측의 정확도를 높이는 것을 중점적으로 고려하여 학습되어야 한다. 즉, 규칙의 일반성을 고려할 필요가 없다. 그러므로 서브그룹 디스커버리가 아닌 분류 및 예측을 위한 규칙 학습 알고리즘을 이용하는 것이 바람직하다. 본 논문에서는 데이터 분석을 위한 서브그룹 디스커버리 알고리즘에 관해 다루고자 한다.

서브그룹 디스커버리를 위한 대표적인 알고리즘으로 MIDOS [1]가 있다. MIDOS는 목표 클래스가 주어졌을 때 조건부가 일정길이 이하의 속성 값의 논리곱으로 표현된 후보 규칙들을 탐색 전략에 따라 평가하여 상위 k개의 규칙을 찾는 알고리즘이다. 평가 척도로는 규칙의 일반성과 특이성을 동시에 평가할 수 있는 가중 상대 정확도(Weighted Relative Accuracy, WRA)[1]를 이용한다.

$$WRA(r)=p(Cond_r)(p(Class|Cond_r)-p(Class)) \quad (1)$$

수식(1) 에서 $p(Cond)$는 규칙의 일반성을 나타내는 것으로 전체 Examples에 대해 규칙 r의 목표 클래스와 관계없이 조건부가 커버하는 Example들의 비율이며 Coverage라고 한다. $p(Class)$는 전체 Examples에 대해 목표 클래스에 해당하는 Example들(이하 Positive example)의 비율이며 데이터 상에서 목표 클래스의 일반적 경향을 나타낸다. $p(Class|Cond_r)$은 규칙이 커버하는 Example들 중에 Positive example의 비율이다. 이 두 가지의 차이는 학습된 서브그룹이 일반적 경향과 얼마나 차이가 나는지, 즉 규칙의 특이성을 보여주는 지표가 되고 Relative Accuracy 혹은 Bias라고 한다. 규칙의 특이성에 일반성을 나타내는 Coverage를 Weight로 삼아 규칙을 평가하는 척도라고 하여 Weighted Relative Accuracy라고 한다.

MIDOS의 경우 정해진 수만큼 규칙을 학습하기 때문에 그 수가 크면 클수록 목표 클래스에 해당하는 데이터들을 많이 커버하는 규칙 집합을 학습하여 다양한 규칙을 학습할 수 있다. 그러나 데이터를 분석하는 사용자의 관점에서 규칙 집합의 크기가 클수록 분석이 어렵다는 단점이 있다.

이러한 단점을 극복하기 위해 [2]에서는 규칙을 분석하는 사용자의 편의를 고려하여 유의미한 규칙을 추려내는 후처리 방안을 제안하고 있다. 후처리 과정은 3단계로 구성되는데, 규칙의 특이성이 높은 흥미로운 규칙을 추려내는 단계, 커버하는 데이터 그룹이 완전히 겹치는 불필요한 규칙을 제거하는 단계, 논리합을 이용하여 두 개 이상의 규칙들을 하나의 규칙으로 표현하여 규칙의 일반성을 높이는 단계이다. 그러나 첫 번째 단계는 사용자가 정해야 할 기준 값들이 많다는 단점이 있다. 또한 세 번째 단계는 규칙의 표현의 한계로 인한 과정이다. 본 논문에서는 이러한 점들을 고려하여 규칙의 표현력을 높이고 개선된 후처리 과정을 제안하고자 한다.

Inputs: dataset D, target t, desired number of rules k, rule length l
Outputs: rule set R
Q = exhaustive set of '$Cond \rightarrow t$' rules with $|Cond| \leq l$
$R = \emptyset$
while $Q \neq \emptyset$ **do**
  Fetch out a rule r from Q according to search strategy
  Compute $WRA(r)$ against D
  if $|R| < k$ then add r to R
  else if $WRA^*(r) \leq \min_{h \in R} WRA(h)$
    then Prune Q by removing all the rules whose conditions are more specific than that of r
  else if $WRA(r) > \min_{h \in R} WRA(h)$
    then Replace the worst element of R with r

Fig. 1. Pseudocode of MIDOS

[2]에서는 또한 후처리 과정 이외에 목표 클래스를 속성 값의 논리곱으로 표현하여 규칙을 학습하는 방안을 제안하였다. 데이터의 한 가지 속성으로 클래스를 분류하여 규칙을 학습하는 것을 넘어 다중 속성들의 속성 값들로 클래스를 분류하여 규칙을 학습할 수 있기 때문에 사용자의 데이터 분석 의도에 부합하는 규칙의 학습이 가능하다. 본 논문에서 또한 결합된 형태의 목표 클래스에 대하여 규칙을 학습하였다.

MIDOS외에 대표적인 서브그룹 디스커버리 알고리즘으로 CN2-SD [3]가 있다. CN2-SD는 규칙을 하나씩 학습하는 과정을 모든 인스턴스들이 커버될 때까지 반복하면서 규칙 집합을 학습한다. 이 알고리즘 또한 규칙을 평가하는 척도로써 WRA를 이용한다. CN2-SD는 분류를 위한 규칙 학습 알고리즘인 CN2 [8], [9]를 바탕으로 서브그룹 디스커버리에 적합하게 개량한 알고리즘이다. CN2 알고리즘은 분류의 정확도를 높이는 것을 중점적으로 규칙을 학습하기 때문에 커버된 인스턴스들은 제외하고 다음 규칙을 학습한다. 반면 CN2-SD는 데이터의 분석 및 해석에 목적을 둔 서술 모형을 학습하기 때문에 인스턴스들에 초기 가중치를 부여하고 이 가중치를 이용하여 WRA를 계산하며 규칙이 학습될 때마다 학습된 규칙에 의해 커버된 인스턴스들의 가중치를 감소시킨다. CN2-SD 방식은 하나의 규칙을 학습하는 방법에 따라 두 가지 종류가 있다. 한 가지는 규칙을 학습하면서 클래스 또한 결정하는 방식 [8], [10]이고 다른 한 가지는 정해진 목표 클래스에 대해서 규칙을 학습하는 방식이다 [4].

모든 인스턴스들을 커버하는 것은 어렵기 때문에 조건을 완화한 변형된 CN2-SD알고리즘들이 있다. 데이터가 커버되었는지 여부와 관계없이 규칙을 학습하는 MIDOS와는 달리 CN2-SD는 커버되지 않은 인스턴스들을 위주로 다음 규칙을 학습하기 때문에 규칙의 다양성이 높고 규칙 집합에 의해서 올바르게 커버되는 인스턴스의 비율이 높다. 그러나 가중치를 이용하여 규칙을 평가하기 때문에 평가 값에 왜곡이 생길 수 있고 따라서 학습된 규칙의 질이 좋지 않을 수 있다는 단점이 있다. 본 논문에서는 이러한 CN2-SD의 단점을 MIDOS 방식의 규칙 학습을 통해 보완하고 CN2-SD의 장점을 살린 서브그룹 디스커버리 알고리즘을 제안한다.

Inputs: dataset D with instance weights initialized to 1, target t
Outputs: rule set R
$R = \emptyset$
while positive examples in D have weight 1 **do**
  $r \leftarrow LearnRuleForClass(D, t)$
  Append r to the end of R
  Decrease the weights of exmpales covered by r

Fig. 2. Pseudocode of CN2-SD(Weighted Covering)

Inputs: dataset D, target t
Outputs: rule r
$b \leftarrow true$
$L \leftarrow$ set of available literals
while not $Homogeneous(D)$ **do**
  $l \leftarrow BestLiteral\ (D, L, t)$
  $b \leftarrow b \wedge l$
  $D \leftarrow \{x \in D \mid x$ is covered by $b\}$
  $L \leftarrow L \setminus \{l' \in L \mid l'$ uses same features by $b\}$
$r \leftarrow$ if b then Class = t

Fig. 3. Pseudocode of CN2-SD(LearnRuleForClass)

서브그룹 디스커버리 알고리즘은 학습된 규칙 집합으로 평가된다. 규칙 집합은 규칙의 복잡성, 일반성, 특이성 등 여러 가지 측면으로 평가할 수 있다 [11]. 분류를 위해 학습된 규칙 집합의 경우 정확도를 평가하는 척도도 있다. 본 논문에서는 규칙 집합의 크기를 통해 규칙 집합의 복잡도를 평가하고 규칙의 커버리지(Coverage), 상대 정확도(Relative Accuracy), 가중 상대 정확도(WRA) 및 규칙 집합이 올바르게 커버하는 인스턴스들의 비율을 나타내는 서포트(Support)를 이용하여 규칙의 일반성 및 특이성을 평가하고자 한다.

III. Internal Disjunctive Rule Learning

이 장에서는 본 논문에서 제안하는 내부 디스정션으로 표현된 규칙을 획득하는 방안을 설명한다. 내부 디스정션으로 표현된 규칙이라는 것은 규칙을 구성하는 하나 이상의 속성들의 각각에 대해 한가지 값만을 갖는 것이 아니라 다수의 값들을 논리합으로 표현한 규칙을 의미한다. 예를 들어 속성 F1, F2에 대해 기존의 규칙은 $F_1=v_1 \wedge F_2=w_1$을 조건부로 가졌다면 내부 디스정션 표현이 가능해짐에 따라 $(F_1=v_1 \vee v_2) \wedge (F_2=w_1)$형식으로 규칙의 조건부를 표현할 수 있다. 한 속성이 하나의 값을 갖는 형태의 논리곱으로 표현하는 기존 방식에 비해 내부 디스정션으로 표현된 규칙은 데이터를 커버하는 영역이 넓어지기 때문에 규칙의 일반성을 증가 시킬 수 있다.

1. 내부 디스정션 규칙 학습을 위한 알고리즘

내부 디스정션으로 표현된 규칙은 서브그룹 디스커버리 알고리즘에 의해 학습된 규칙들을 대상으로 내부 디스정션 알고리즘에 의해 획득할 수 있다. 본 논문에서 제안하는 방안은 규칙의 조건부를 구성하는 속성 각 각에 대해서 독립적으로 내부 디스정션 과정을 거친다. 즉 n개의 속성으로 구성된 규칙 하나는 최대 n개의 일반화된 규칙이 생성된다. 모든 속성에 대해 내부 디스정션 표현이 가능하도록 했을 경우 규칙에 대한 해석이 어렵기 때문에 하나의 속성에 대한 내부 디스정션으로 제한하였다.

내부 디스정션 알고리즘에서 속성 값을 하나씩 추가하며 규칙을 평가하는데 이 때 WRA만으로 규칙을 평가하는 것이 아니라 편향성(Bias)를 고려한다. 내부 디스정션 표현은 규칙의 일반성, 즉 커버리지를 증가시키는 효과가 있다. 따라서 편향성를 고려하지 않고 WRA만으로 규칙을 평가하면 특이성이 높은 규칙은 살아남지 못하고 일반성만 높은 규칙들이 남게 된다. 따라서 규칙의 일반성과 특이성을 모두 고려하기 위해 편향성과 WRA가 모두 향상될 때까지 속성 값을 추가한다.

내부 디스정션 알고리즘을 이용하면 규칙의 표현력을 증대시킬 수 있지만 조건부를 구성하는 속성 수만큼 규칙을 추가 생성하기 때문에 규칙의 수 또한 증가하여 규칙 집합의 분석이 어렵다는 문제점이 있다. 이것을 해결하기 위해 다음 절에서 규칙 집합을 간소화 하는 방안을 제안한다.

2. 규칙 집합의 간소화

2.1 규칙 선별 방안

내부 디스정션 알고리즘에 의해 커진 규칙 집합의 크기를 줄이기 위해 흥미로운 규칙들을 추려내기 위한 방안이 필요하다. 가장 기본적으로 고려할 수 있는 방안은 규칙을 평가하는 척도인 WRA의 평균을 기준으로 평균 이상인 규칙들만 남기는 것이다. WRA의 평균을 기준으로 규칙을 선별하는 방안은 규칙의 커버리지와 편향성를 모두 고려하여 질이 좋은 규칙을 선별하는 것을 의미한다. 규칙의 특이성을 중점적으로 고려하기 위해서는 규칙 집합을 구성하는 규칙들의 편향성 평균을 기준으로 규칙을 선별할 수 있다.

그러나 앞의 두 방안은 규칙들이 커버하는 학습 데이터를 고려하지 않기 때문에 커버하는 데이터가 유사한 규칙들이 선택되는 문제점이 있다. 커버하는 인스턴스들이 다른 규칙들을 선택하기 위한 관점에서는 규칙들 간의 커버하는 인스턴스들이 얼마나 비슷한지를 나타내는 유사도를 고려할 필요가 있다. 본 논문에서 정의한 두 규칙 간의 유사도는 다음과 같다.

$$Similarity(r_i,r_j)=\frac{Pos(Cond_{ri} \wedge Cond_{rj})}{Pos(Cond_{ri})} \qquad (2)$$

수식(2)는 두 규칙 r_i, r_j에 대해 r_j에 대한 r_i의 유사도를 나타낸다. $Pos(Cond_{ri})$ 는 규칙 r_i에 의해 올바르게 커버된 인스턴스의 수를 나타내고 $Pos(Cond_{Ri} \wedge Cond_{Rj})$ 는 두 규칙이 동시에 올바르게 커버하는 인스턴스의 수를 나타낸다. 즉, 자신이 올바르게 커버한 인스턴스들 중 다른 규칙이 커버하지 않은 인스턴스를 많이 포함하고 있다면 유사도는 낮아지게 된다.

앞에서 정의한 규칙의 유사도를 기반으로 WRA를 재평가한다. 이 평가 값은 규칙 r에 대한 WRA를 규칙 집합을 구성하는 모든 규칙들 각 각에 대한 유사도의 총합으로 나눈 것으로 정의되고 sharedWRA라고 명명한다.

$$sharedWRA(r)=\frac{WRA(r)}{\sum_{r'} similarity(r,r')} \qquad (3)$$

커버하는 영역이 서로 다른 규칙들을 선별하기 위해 규칙 집합을 구성하는 규칙들의 sharedWRA의 평균을 기준으로 평균 이상인 규칙들은 선별한다.

규칙 집합을 간소화하기 위해 제안한 규칙 선별 방안들은 특정 척도에 대해 평균값을 기준으로 삼고 있다. 특정 값을 기준으로 선별할 수 있지만 파라미터 조정이 필요하기 때문에 편의상 절반을 제거할 목적으로 평균값을 이용하였다.

2.2 불규칙한 규칙 제거

내부 디스정션 과정은 규칙 집합의 각 규칙에 대해 독립적으

로 진행되므로 커버하는 영역이 다른 규칙이 커버하는 영역에 완전히 포함되는 규칙이 존재할 수 있다. 따라서 이러한 불필요한 규칙들을 제거할 필요가 있다. 불필요한 규칙의 정의는 다음과 같다.

F_R은 규칙 R에 대해 규칙의 조건부를 구성하는 속성 집합이고 V_{Rf}는 속성 f에 대해 규칙 R의 조건부를 구성하는 속성 f의 값들의 집합이라고 정의하자. 임의의 두 규칙 A, B에 대해 $F_A \subseteq F_B$ 일 때 임의의 속성 $f \in F_A$에 대해 $V_{Bf} \subseteq V_{Af}$ 이고, $WRA(B) \leq WRA(A)$ 이면 규칙 B는 규칙 A에 관해 불필요한 규칙이다.

커버하는 영역이 더 넓은 규칙을 선별하면서도 특이성이 높은 규칙을 놓치지 않기 위해 WRA를 고려하였다. 커버하는 영역이 다른 규칙에 완전히 포함되면서도 WRA가 더 높다는 것은 일반성을 능가할 정도로 특이성이 높은 규칙임을 의미한다.

IV. Proposed Subgroup Discovery Algorithm

앞에서 언급했듯이 MIDOS와 CN2-SD는 각 각의 장점이 있지만 한계점이 있다. MIDOS는 정해진 수만큼 규칙을 학습하는데 커버하는 영역이 유사한 규칙들이 많아 규칙의 다양성이 부족한 문제가 있다. 학습하는 규칙의 수를 크게 하면 다양성이 증가하여 올바르게 커버된 인스턴스의 비율이 증가하긴 하지만 규칙 집합의 크기가 클수록 분석이 어렵다. 규칙 집합의 크기를 줄이기 위한 후처리 방안을 제안한 연구가 있었지만 그 과정이 복잡하고 사용자 파라미터가 많다는 문제가 있다.

CN2-SD는 목표 클래스에 해당하는 인스턴스들을 일정 비율 이상으로 커버하는 규칙 집합을 찾기 위해 커버되지 않은 인스턴스를 중점적으로 커버하는 규칙을 찾는다. 따라서 데이터를 다각도로 조명한 규칙의 학습이 가능하여 규칙의 다양성이 높다. 그러나 인스턴스의 가중치를 이용하여 규칙을 평가하기 때문에 평가에 왜곡이 생겨 개개의 규칙의 질을 보장하지 못한다는 한계가 있다.

본 논문에서는 MIDOS와 CN2-SD 각 각의 장점을 살려서 단점을 보완하고 내부 디스정션 표현이 가능한 규칙을 학습하는 CN2-SD+MIDOS-ID 알고리즘을 제안한다. 기존의 CN2-SD 알고리즘이 규칙을 하나씩 학습하는 과정을 반복하였다면 제안하는 알고리즘은 다수의 규칙을 학습하는 과정을 반복하는 것이 대표적인 특징이다. 즉, 그림 1에서 보인 CN2-SD의 가중 커버링 알고리즘(Weighted Covering algorithm)에서 규칙 학습하는 방안을 MIDOS 방식으로 변형한 것이다. MIDOS 알고리즘에 의해 학습된 규칙 집합은 내부 디스정션으로 표현된 규칙 집합으로 확장되고 규칙 집합 간소화 과정을 거치게 된다. 남은 규칙들에 의해 커버된 인스턴스들의 가중치를 낮춘다.

인스턴스의 가중치를 업데이트 하는 방식에는 [3]에서 두 가지 방안을 제안하고 비교하고 있다. 본 논문에서는 두 가지 중 성능이 우수했던 추가 가중치(Additive Weights) 방식으로 업데이트한다. 추가 가중치 방식은 인스턴스를 커버하는 규칙의 수에 1을 더한 값의 역수로 가중치를 결정한다. 예를 들어 초기 가중치를 1로 부여하고 하나의 규칙에 의해 커버된 인스턴스의 가중치는 2분의 1이 된다.

이 가중치를 이용하여 WRA를 계산하는 경우 실제로는 상대 정확도가 음수인, 즉 목표 클래스의 일반적 경향성에도 미치지 못하는 규칙이 학습되는 경우가 발생한다. 목표 클래스에 대한 특성을 분석하기 위한 규칙을 학습하고자 할 때, 실제 WRA가 음수인 규칙은 아무런 의미가 없기 때문에 이러한 규칙들을 제거하는 과정이 그림 4에 나타난다. 남은 규칙이 없는 경우 목표 클래스의 인스터스들의 커버 여부와 관계없이 규칙 학습을 멈추고 지금까지 학습한 규칙 집합을 반환한다.

```
Inputs: dataset D with instance weights initialized to 1,
        target t, desired number of rules k, rule length l
Outputs: rule set R
R = Ø
while positive examples in D have weight 1 do
   R'← MIDOS(D, t, k, l)
   InternealDisjunction(R')
   Remove meaningless rules of which real WRA<0
   if |R'| is zero then break;
   Select Interesting rules (by Bias, WRA or sharedWRA)
   Remove Redundant rules
   Append R' to the end of R
   Decrease the weights of examples covered by R'
```

Fig. 4. Pseudocode of CN2-SD+MIDOS-ID

V. Experimental Results

1. 실험 환경

1.1 데이터

부산시 교통사고 데이터를 이용하여 실험하였다. 이 데이터는 2011년 1월부터 2014년 8월까지 총 44개월 동안 발생한 50,709건의 교통사고 기록으로 구성되어 있다.

시간대에 따른 음주운전 사고의 특성을 분석하기 위해 'Time'과 'Drunk driving' 속성 값을 조합한 목표 클래스에 대해 규칙을 학습하였다. 즉, 각 시간대별 음주운전 사고의 특성을 규칙으로 학습하였다.

1.2 평가 척도

규칙 집합은 다섯 가지의 척도를 이용하여 평가된다

[3].Table 3에서는 다섯 가지 척도를 보여주고 있다. 커버리지(Coverage)는 규칙 집합을 구성하는 각 각의 규칙들의 커버리지를 평균한 값이다. 편향성(Bias) 또한 각 각의 규칙들의 편향성을 평균한 값이다. 서포트(Support)는 전체 목표 클래스의 인스턴스 중 규칙 집합에 의해서 커버된 인스턴스의 비율을 의미한다. 하나의 인스턴스가 여러 개의 규칙에 의해 커버되더라도 한 번만 세어진다. 크기(Size)는 규칙 집합을 구성하는 규칙의 개수를 나타내며 특이성(Unusualness)은 평균 WRA이다.

Table 2. Attributes of traffic accident data

Attribute	Value
Day type	Weekday, Weekend
Time	00:00~08:00, 08:00~13:00, 13:00~17:00, 17:00~21:00, 21:00~24:00
Place	20 categories (Intersection, Crosswalk, Motorway, Schoolzone, Marketpalce, Parking area, etc.)
Vehicle type	12 types (Subcompact, Compact, Mid-sized, Full-sized, SUV, Truck, Bus, Motorcycle, etc.)
Drunk Driving	True, False
Precipitation	None, 0~1mm, 1~5mm, 5~20mm, 20~80mm, 80~150mm, 180mm+
Temperature	Below zero, 0~10°C, 10~20°C, 20°C+

Table 3. Classes targeted in Experiment

Target Class
Time = 00:00~08:00 ∧ Drunk driving = True
Time = 08:00~13:00 ∧ Drunk driving = True
Time = 13:00~17:00 ∧ Drunk driving = True
Time = 17:00~21:00 ∧ Drunk driving = True
Time = 21:00~24:00 ∧ Drunk driving = True

Table 4. Evaluation measures of rule set

Measure	Computation
Coverage	$COV = \frac{1}{n_R}\sum_{i=1}^{n_R} Cov(R_i) = \frac{1}{n_R}\sum_{i=1}^{n_R}\frac{n(Cond_i)}{N}$
Bias	$BIAS = \frac{1}{n_R}\sum_{i=1}^{n_R} Bias(R_i)$
Support	$SUP = \frac{1}{Pos} Pos(Class \cdot \bigvee_{Class \leftarrow Cond_i} Cond_i)$
Size	$SIZE = n_R$
Unusualness	$WRACC = \frac{1}{n_R}\sum_{i=1}^{n_R} WRA(R_i)$

2. 실험 결과 및 분석

2.1 내부 디스정션 표현의 효과

내부 디스정션 표현의 효과를 확인하기 위해 기존 서브그룹 디스커버리 알고리즘으로부터 학습한 규칙 집합과 그 규칙 집합으로부터 내부 디스정션 과정과 불필요한 규칙 제거 과정을 거쳐서 얻은 규칙 집합간의 성능을 비교하였다. 윌콕슨 부호-순위 검정 방식 [13]을 이용하여 신뢰도 95% 수준으로 결과를 검정하였다.

Table 4 는 CN2-SD로 학습한 규칙 집합에 대한 비교이다. CN2-SD-ID는 내부 디스정션 및 불필요한 규칙 제거 과정을 거치는 방법을 의미한다. 기존 규칙 집합에 비해 내부 디스정션 과정을 거친 규칙 집합의 커버리지와 서포트가 증가한 것을 확인할 수 있다. 특히 21시~24시의 음주운전 사고에 대한 규칙 집합은 CN2-SD-ID가 CN2-SD에 비해 더 적은 수의 규칙으로 많은 목표 클래스의 인스터스들을 커버하는 것을 확인할 수 있다.

Table 5. Performance Comparison of CN2-SD and CN2-SD involving internal disjunction process

Time	Method	Cov.	Bias	Sup.	Size	Unusual.
00~08	CN2-SD	0.149	0.0162	0.963	9	0.00181
	CN2-SD-ID	**0.233**	0.0180	**0.977**	11	**0.00282**
08~13	CN2-SD	0.127	0.0039	0.610	4	0.00039
	CN2-SD-ID	**0.153**	0.0031	**0.752**	5	0.00032
13~17	CN2-SD	0.178	0.0030	0.839	5	0.00051
	CN2-SD-ID	0.178	0.0030	0.839	5	0.00051
17~21	CN2-SD	0.284	0.0039	0.372	1	0.00112
	CN2-SD-ID	0.284	0.0039	0.372	1	0.00112
21~24	CN2-SD	0.132	0.0092	0.894	9	0.00080
	CN2-SD-ID	**0.162**	0.0106	**0.914**	**8**	**0.00109**
Wilcoxon signed-rank test		**0**	3	**0**	3	3

Table 5 는 MIDOS에서 내부 디스정션 및 불필요한 규칙 제거 과정을 거친 효과를 확인한 결과이다. MIDOS 알고리즘으로부터 학습할 규칙의 수 k값을 10으로 설정하여 분석한 결과이다. MIDOS에서 역시 커버리지와 서포트가 증가하는 것을 확인할 수 있으며 개수가 고정된 MIDOS에 비해 규칙 집합이 컴팩트하면서 성능이 우수한 것을 확인할 수 있다.

Table 6. Performance Comparison of MIDOS(k=10) and MIDOS(k=10) involving internal disjunction process

Time	Method	Cov.	Bias	Sup.	Size	Unusual.
00~08	MIDOS	0.142	0.0279	0.870	10	0.00227
	MIDOS-ID	**0.397**	0.0141	**0.898**	**3**	**0.00518**
08~13	MIDOS	0.084	0.0041	0.610	10	0.00028
	MIDOS-ID	**0.169**	0.0038	0.610	**3**	**0.00052**
13~17	MIDOS	0.129	0.0058	0.798	10	0.00047
	MIDOS-ID	**0.269**	0.0032	0.798	**3**	**0.00081**
17~21	MIDOS	0.105	0.0088	0.712	10	0.00054
	MIDOS-ID	**0.159**	0.0086	**0.718**	**6**	**0.00086**
21~24	MIDOS	0.179	0.0109	0.894	10	0.00105
	MIDOS-ID	**0.332**	0.0082	**0.901**	**5**	**0.00166**
Wilcoxon signed-rank test		**0**	0	**0**	**0**	**0**

Table 7. Performance comparision of rule selection methods

Time	Method	Cov.	Bias	Sup.	Size	Unusual.
00~08	MIDOS-ID-*half*(Bias)	0.079	**0.0385**	0.273	2	0.00300
	MIDOS-ID-*half*(WRA)	0.397	0.0141	**0.898**	3	0.00518
	MIDOS-ID-*half*(SharedWRA)	0.397	0.0141	**0.898**	3	0.00518
08~13	MIDOS-ID-*half*(Bias)	0.057	**0.0060**	0.248	3	0.00033
	MIDOS-ID-*half*(WRA)	0.232	0.0031	**0.573**	2	0.00066
	MIDOS-ID-*half*(SharedWRA)	0.232	0.0031	**0.573**	2	0.00066
13~17	MIDOS-ID-*half*(Bias)	0.069	**0.0090**	0.397	3	0.00060
	MIDOS-ID-*half*(WRA)	0.269	0.0032	**0.798**	3	0.00081
	MIDOS-ID-*half*(SharedWRA)	0.269	0.0032	**0.798**	3	0.00081
17~21	MIDOS-ID-*half*(Bias)	0.047	**0.0164**	0.205	2	0.00081
	MIDOS-ID-*half*(WRA)	0.156	0.0090	0.571	4	0.00101
	MIDOS-ID-*half*(SharedWRA)	0.183	0.0077	**0.705**	5	0.00093
21~24	MIDOS-ID-*half*(Bias)	0.139	**0.0153**	0.409	2	0.00204
	MIDOS-ID-*half*(WRA)	0.218	0.0119	0.683	3	0.00200
	MIDOS-ID-*half*(SharedWRA)	0.343	0.0094	**0.864**	4	0.00182

2.2 규칙 선별 방안의 비교

본 논문에서는 내부 디스정션 과정에 의해 커진 규칙 집합의 크기를 축소시키기 위해 규칙 선별 방안을 고안하였다. 규칙간에 중복으로 커버하는 데이터의 양을 기반으로 수식(3)과 같이 유사도를 정의하고 WRA가 크면서 유사성이 낮은 규칙을 선별하는 방안이다.

표 6은 제안한 규칙 선별 방안과 기본적으로 고려할 수 있는 WRA 평균을 기준으로 선별하는 방안 및 편향성 평균을 기준으로 선별하는 방안에 대해 각 각 학습한 규칙 집합의 성능을 비교한 결과이다.

편향성 평균 기준의 규칙 선별 방안은 단연 편향성 측면에서의 결과가 높게 나타났다. 다만 규칙의 특이성을 중점적으로 규칙을 선별하므로 커버리지와 서포트가 낮다. sharedWRA 평균 기준의 규칙 선별 방안은 다른 방안에 비해 안정적으로 높은 서포트를 유지하는 것을 확인할 수 있다. 유사도를 고려하여 규칙을 선별한 sharedWRA 평균 기준 방안이 WRA평균 기준 방안에 비해 1개의 규칙을 더 선별하면서 서포트를 높이는 것을 확인할 수 있다.

그러나 규칙 선별 방안은 특정 방안이 확연히 뛰어나다고 단정할 수 없고 데이터를 분석하는 사용자의 의도에 적합한 방식을 택하는 것이 바람직할 것으로 보인다. 일반적 경향과 상당히 다른 특이한 규칙을 찾는 것이 주된 목적이라면 편향성을 기준으로 규칙을 선별하고 목표 클래스에 해당하는 데이터를 다각도로 최대한 설명하는 규칙 집합을 학습하는 것이 목적이라면 안정적으로 높은 서포트를 유지할 수 있는 SharedWRA 평균 기준 규칙 선별 방안을 이용할 수 있다.

2.3 제안하는 서브그룹 디스커버리 알고리즘의 성능

표 7은 본 논문에서 제안한 서브그룹 디스커버리 방식의 성능을 CN2-SD-ID방식과 비교하여 분석한 결과이다. 99%의 목표 클래스의 인스터스를 커버할 때까지 규칙을 학습하였다. 라운드(Round)는 규칙 집합 학습을 완료할 때까지 규칙 집합 도출을 반복한 횟수를 나타낸다.

규칙을 하나씩 학습하는 CN2-SD 알고리즘에 비해 한 라운드에서 다수의 규칙을 학습하는 제안 방안이 적은 반복 횟수만으로 안정적으로 서포트를 만족시키는 것을 확인할 수 있다. 역시 편향성 평균을 기준으로 규칙을 선별하며 학습하는 방안이 다른 방안들에 비해 편향성 측면에서 우수하지만 서포트를 만족시키지 못하거나 만족시키기 위해서는 규칙 집합의 사이즈가 상당히 커지는 경향을 보인다. 또한 k값이 증가할수록 *half*(sharedWRA) 규칙 선별 방안이 크기가 작으면서 서포트와 특이성 측면에서 우수한 경향을 확인할 수 있다.

MIDOS와 제안방안을 비교했을 때 MIDOS 알고리즘만으로는 커버할 수 없었던 목표 클래스의 인스턴스들을 커버하는 규칙을 학습할 수 있음으로 인해 서포트가 증가하는 것을 확인할 수 있다. 즉, MIDOS로는 학습할 수 없었던 새로운 측면의 규칙

이 학습되면서 규칙의 다양성을 증가시킬 수 있다. 이것은 규칙의 정량적 평가보다 정성적 측면에서 평가가 필요한 부분이다.

표 8은 Time = 00:00~08:00 ∧ Drunk driving = True를 목표 클래스로 학습한 규칙들을 학습 알고리즘에 따라 나누어 보여준다. 새벽시간 대에 발생한 음주운전의 특징으로 세 가지 방안 모두 차종에 대한 특징을 많이 보여주고 있다. 그리고 제안한 알고리즘을 통해 기존 알고리즘으로는 발견하지 못했던 사고 장소 유형의 특징과 강수량에 대한 특징을 찾아내고 있다.

Table 8. Performance of proposed method

Time	Method	Cov.	Bias	Sup.	Size	Unusual.	Round
00~08	CN2-SD-ID	0.233	0.0180	0.977	11	0.00282	15
	CN2-SD+MIDOS-ID	0.290	0.0158	**0.995**	16	0.00324	7
	CN2-SD+MIDOS-ID -*half*(Bias)	0.138	**0.0213**	0.994	33	0.00214	16
	CN2-SD+MIDOS-ID -*half*(SharedWRA)	**0.327**	0.0126	**0.995**	**14**	**0.00355**	**7**
08~13	CN2-SD-ID	0.153	0.0031	0.752	5	0.00032	6
	CN2-SD+MIDOS-ID	0.155	0.0029	**0.991**	23	0.00022	**6**
	CN2-SD+MIDOS-ID -*half*(Bias)	0.073	**0.0036**	**0.991**	57	0.00017	27
	CN2-SD+MIDOS-ID -*half*(SharedWRA)	**0.164**	0.0026	**0.991**	**19**	**0.00027**	7
13~17	CN2-SD-ID	0.178	0.0030	0.839	5	0.00051	8
	CN2-SD+MIDOS-ID	0.167	0.0034	**0.996**	25	0.00046	9
	CN2-SD+MIDOS-ID -*half*(Bias)	0.102	**0.0046**	0.993	60	0.00035	36
	CN2-SD+MIDOS-ID -*half*(SharedWRA)	**0.203**	0.0033	**0.996**	**19**	**0.00053**	**9**
17~21	CN2-SD-ID	0.284	0.0039	0.372	1	0.00112	2
	CN2-SD+MIDOS-ID	0.081	0.0084	**0.946**	35	0.00041	15
	CN2-SD+MIDOS-ID -*half*(Bias)	0.048	0.0081	0.927	36	0.00032	23
	CN2-SD+MIDOS-ID -*half*(SharedWRA)	0.087	0.0090	0.930	**21**	0.00047	12
21~24	CN2-SD-ID	0.162	0.0106	0.914	8	0.00109	9
	CN2-SD+MIDOS-ID	**0.300**	0.0064	**0.996**	**10**	0.00131	**4**
	CN2-SD+MIDOS-ID -*half*(Bias)	0.109	**0.0106**	0.992	38	0.00074	25
	CN2-SD+MIDOS-ID -*half*(SharedWRA)	0.259	0.0076	0.995	10	**0.00145**	**4**

Wilcoxon Test	CN2-SD+MIDOS-ID	CN2-SD+MIDOS-ID -*half*(bias)	CN2-SD+MIDOS-ID -*half*(sharedWRA)
Coverage	7	0	5
Bias	7	**0**	5
Support	**0**	**0**	**0**
Size	0	0	0
Unusualness	7	0	6

Table 10. Analysis of rule sets

Subgroup	Cov.	Bias	WRA
CN2-SD-ID			
Day type=Weekend	0.284	0.0213	0.00605
Vehicle type=Sports|Compact|Subcompact|Mid-sized|Full-sized	0.565	0.0090	0.00507
Temperature=10~20℃|Below zero	0.335	0.0104	0.00349
Vehicle type=Sports|Compact|Mid-sized|Full-sized	0.508	0.0056	0.00284
Vehicle type=Sports|Subcompact|Motorcycle	0.125	0.0193	0.00242
Vehicle type=Sports|Subcompact	0.058	0.0400	0.00232
Vehicle type=Subcompact	0.057	0.0394	0.00223
Vehicle type=SUV|Sports|Motorcycle	0.163	0.0116	0.00189
Vehicle type=SUV|Sports	0.095	0.0187	0.00178
Vehicle type=SUV	0.094	0.0181	0.00170
Temperature=0~10℃|Below zero	0.280	0.0045	0.00127
MIDOS-ID-*half*(sharedWRA)			
Day type=Weekend	0.284	0.0213	0.00605
Vehicle type=SUV|Sports|Compact|Subcompact|Mid-sized	0.573	0.0105	0.00601
Temperature=10~20℃|Below zero	0.335	0.0104	0.00349
CN2-SD+MIDOS-ID-*half*(sharedWRA)			
Day type=Weekend	0.284	0.0213	0.00605
Vehicle type=SUV|Sports|Compact|Subcompact|Mid-sized	0.573	0.0105	0.00601
Vehicle type=SUV|Sports|Compact|Subcompact|Full-sized	0.295	0.0191	0.00564
Vehicle type=Sports|Compact|Subcompact|Mid-sized|Full-sized	0.565	0.0090	0.00507
Vehicle type=SUV|Sports|Subcompact|Full-sized|Motorcycle	0.305	0.0160	0.00487
Vehicle type=Sports|Compact|Subcompact|Mid-sized|Full-sized|Van	0.611	0.0071	0.00436
Vehicle type=SUV|Sports|Subcompact	0.152	0.0264	0.00401
Vehicle type=SUV|Sports|Compact|Subcompact|Mid-sized|Motorcycle|Truck	0.748	0.0049	0.00368
Temperature=10~20℃|Below zero	0.335	0.0104	0.00349
Vehicle type=Sports|Subcompact|Motorcycle	0.125	0.0193	0.00242
Place=Highway|Intersection|Tunnel	0.158	0.0119	0.00188
Temperature=0~10℃|Below zero	0.280	0.0045	0.00127
Day type=Weekday, Vehicle type=SUV|Motorcycle	0.112	0.0056	0.00063
Precipitation=0~1mm|1~5mm, Temperature=20℃+	0.034	0.0100	0.00034

VI. Conclusion

서브그룹 디스커버리는 다양한 속성으로 구성된 데이터로부터 정해진 클래스에 대해 유의미한 규칙을 발굴하는 지도 학습 기법이다. 규칙은 일반적 경향과 상당히 다른 경향을 보일 뿐만 아니라 일반성이 높을수록 우수하다. 규칙의 조건부를 내부 디스정션 형식으로 표현하여 많은 데이터를 커버하는 규칙을 학습할 수 있다. 본 논문에서 제안한 내부 디스정션 규칙 획득 방안은 규칙 집합의 크기가 증가하기 때문에 사용자의 의도에 맞는 규칙 선별 방안이 필요하다. 또한 불필요한 규칙의 제거를 통해 의미있는 규칙을 선별할 수 있다.

그럼에도 불구하고 MIDOS 방식의 규칙 학습은 목표 클래스 데이터를 상당히 커버하지 못하고 CN2-SD만큼의 흥미로운 규칙을 발견하지 못하는 문제가 있다. 또한 CN2-SD는 무의미한 규칙을 학습하여 결국 데이터를 사용자가 원하는 수준만큼 커버하지 못하는 문제가 있다. 이러한 점들을 극복하고 각 각의 장점을 수용할 수 있는 서브그룹 디스커버리 알고리즘을 제안하였다. 기존의 방식에 비해 안정적으로 규칙을 학습하면서도 특이성이 높은 규칙 집합을 학습할 수 있다.

REFERENCES

[1] S. Wrobel, “An algorithm for multi-relational discovery of subgroups,” Principles of Data Mining and Knowledge Discovery, vol. 1763, pp. 78-87, 1997.

[2] J. Kim and K. R. Ryu, “Mining Traffic Accident Data by Subgroup Discovery Using Combinatorial Targets,” Computer Systems and Applications (AICCSA), pp. 1-6, Nov. 2015.

[3] N. Lavrač, B. Kavšek, P. Flach and L. Todorovski, “Subgroup Discovery with CN2-SD,” The Journal of Machine Learning Research, vol. 5, pp. 153-188, Dec. 2004.

[4] P. Flach, “Machine Learning: The Art and Science of Algorithms that Make Sense of Data,” Cambridge University Press, 2012.

[5] G. Wets, K. Vanhoof, B. Depaire, “Traffic accident segmentation by means of latent class clustering,” Accident Analysis & Prevention, vol. 40, No. 4, pp. 1257-1266, July 2008.

[6] J. Kim and K. R. Ryu, “Comparison of Association Rule Learning and Subgroup Discovery for Mining Traffic Accident Data,” Journal of Intelligence and Information Systems, vol. 21, No. 4, pp. 1-16, Dec. 2015.

[7] B. Kavšeka and N. Lavrač, “APRIORI-SD: ADAPTING ASSOCIATION RULE LEARNING TO SUBGROUP DISCOVERY,” Applied Artificial Intelligence: An International Journal, vol. 20, No. 7, pp. 543-583, 2006.

[8] M. J. del Jesus, P. González, F. Herrera and M. Mesonero, “Evolutionary fuzzy rule induction process for subgroup discovery: a case study in marketing,” IEEE Transactions on Fuzzy Systems, vol. 15, No. 4, pp. 578-592, Aug. 2007.

[9] P. Clark and T. Niblett, “The CN2 induction algorithm,” Machine Learning, vol. 3, No. 4, pp. 261-283, Mar. 1989.

[10] P. Clark and R. Boswell, “Rule Induction with CN2: Some Recent Improvements,” Machine Learning — EWSL-91, vol. 482, pp. 151-163, Mar. 1991.

[11] W. W. Cohen, “Fast Effective Rule Induction,” Proceedings of the twelfth international conference on machine learning, pp. 115-123, July 1995.

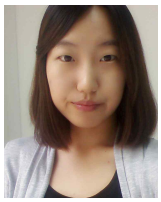
[12] F. Herrera, C. J. Carmona, P. González and M. J. Del Jesus, “An overview on subgroup discovery: foundations and applications,” Knowledge and information systems, vol. 29, No. 3, pp. 495-525, 2011.

[13] F. Wilcoxon, “Some rapid approximate statistical procedures,” Annals of the New York Academy of Sciences, vol. 52, No. 6, pp. 808-814, Mar. 1950.

[14] H. Song, M. Kull, P. Flack and G. Kalogridis, “Subgroup Discovery with Proper Scoring Rules,” Joint European Conference on Machine Learning and Knowledge Discovery in Databases, vol. 9852 pp. 492-510, Sept. 2016.

[15] S. Sumyea, “Subgroup Discovery Algorithms: A Survey and Empirical Evaluation,” Journal of Computer Science and Technology, vol. 31, no. 3, pp. 561-576, May, 2016

Authors



Seyoung Kim received the B.S. degrees in Mathematics and M.S. degrees in Computer Science and Engineering from Pusan National University, Korea, in 2014 and 2016, respectively.

She is currently a Ph.D candidate in the Department of Computer Science and Engineering at Pusan National University, Busan, Korea. She is interested in artificial intelligence, machine learning and data mining.

Kwang Ryel Ryu received the B.S. and M.S. degrees in Electronic Engineering from Seoul National University, Korea, in 1979 and 1981, respectively, and received the Ph.D. degrees in Computer Science and Engineering from the University of Michigan, U.S, in 1992. Dr. Ryu joined the faculty of the Department of Computer Science and Engineering at Pusan National University, Busan, Korea, in 1993. He is currently a Professor in the Department of Computer Science and Engineering, Pusan National University. He is interested in artificial intelligence, machine learning and data mining.