ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2015 December: 21(4): 1~16 http://dx.doi.org/10.13088/jiis.2015.21.4.001

교통사고 데이터의 마이닝을 위한 연관규칙 학습기법과 서브그룹 발견기법의 비교*

김정민
부산대학교 전자전기컴퓨터공학과
(jeongminkim.islab@gmail.com)

류광렬
부산대학교 전자전기텀퓨터공학과
(krryu@pusan.ac.kr)

교통사고의 원인을 규명하고 미래의 사고를 방지하기 위한 노력의 일환으로 데이터 마이닝 기법을 이용한 교통 데이터 분석의 연구가 이루어지고 있다. 하지만 기존의 교통 데이터를 이용한 마이닝 연구들은 학습된 결과를 사람이 이해하기 어려워 분석에 많은 노력이 필요하다는 문제가 있었다. 본 논문에서는 많은 속성들로 표현된 교통사고 데이터로부터 유용한 패턴을 발견하기 위해 규칙 학습 기반의 데이터 마이닝 기법인 연관규칙 학습기법과 서브그룹 발견기법을 적용하였다. 연관규칙 학습기법은 비지도 학습 기법의 하나로 데이터 내에서 동시에 많이 등장하는 아이템(item)들을 찾아 규칙의 형태로 가공해 주며, 서브그룹 발견기법은 사용자가 지정한 대상 속성이 결론부에 나타나는 규칙을 학습하는 지도학습 기반 기법으로 일반성과 흥미도가 높은 규칙을 학습한다. 규칙 학습 시 사용자의 의도를 반영하기 위해서는 하나 이상의 관심 속성들을 조합한 합성 속성을 만들어 규칙을 학습할 수 있다. 규칙이 도출되고 나면 후처리 과정을 통해 중복된 규칙을 제거하고 유사한 규칙을 일반화하여 규칙들을 더 단순하고 이해하기 쉬운 형태로 가공한다. 교통사고 데이터를 대상으로 두 기법을 적용한 결과 대상 속성을 지정하지 않고 연관규칙 학습기법을 적용하는 경우 사용자가 쉽게 알기 어려운 속성 사이의 숨겨진 관계를 발견할 수 있었으며, 대상 속성을 지정하여 연관규칙 학습기법과 서브그룹 발견기법을 적용하는 경우 파라미터 조정에 많은 노력을 기울여야 하는 연관규칙 학습기법에 비해 서브그룹 발견기법이 흥미로운 규칙들을 더 쉽게 찾을 수 있음을 확인하였다.

주제어 : 데이터 마이닝, 연관규칙 학습, 서브그룹 발견, 교통사고 데이터, 규칙 학습

논문접수일 : 2015년 9월 27일 논문수정일 : 2015년 12월 7일 게재확정일 : 2015년 12월 8일
교신저자 : 류광렬

1. 개요

교통사고 데이터는 사고 당시 상황을 기록한 데이터로 운전자, 사고 장소, 사고 종류, 차량 정보, 법규 위반 사항 등 다양한 속성으로 이루어져 있다. 데이터 마이닝은 이러한 데이터로부터 잘 알려져 있지 않은 유용한 패턴을 찾는 과정이다(Witten et al., 2011). 패턴은 규칙, 의사결정 트리, 특정 데이터의 집합과 같이 여러 가지 형태로 표현될 수 있으며 데이터의 특성이나 속성들 사이의 관계에 대한 정보를 제공한다. 어떤 패턴이 유용하다는 것은 패턴에 담긴 정보가 새롭고 향후에 유용하게 활용 가능하다는 것을 뜻한다. 예를 들어 'IF *Drunk Driving* = *true* THEN *Driving Career* ≤ *7 years*' 와 같은 규칙은 그 동안 잘 몰랐던 음주운전 사고와 운전 경력의 관계

* 이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음.
본 논문은 BK21플러스, IT기반 융합산업 창의인력사업단에 의하여 지원되었음.

를 알려주는 것으로 음주운전 예방 교육 프로그램 개발에 활용할 수 있다.

패턴을 발견하기 위한 데이터 마이닝 기법에는 크게 군집화, 서브그룹 발견기법, 연관규칙 학습기법의 세 가지가 있다. 군집화 기법의 경우 주어진 데이터를 서로 유사한 몇 개의 그룹으로 묶어 주지만 나누어진 그룹 자체로부터 명시적인 지식을 얻을 수 없고, 목적에 맞는 지식 발견을 위해서는 후속 분석 작업이 수반되어야 하는 문제가 있다 (Depaire, 2008). 반면 연관규칙 학습기법과 서브그룹 발견기법의 경우 데이터 내에 존재하는 중요한 경향들을 'IF *Cond* THEN *Class*'의 규칙 형태로 제공한다. *Cond*과 *Class*는 속성에 값이 지정된 리터럴(literal)의 조합으로 구성되어 있다. 따라서 규칙의 의미를 이해하기 쉽고 규칙에 대한 평가 함수가 존재하기 때문에 각 규칙의 중요성을 객관적으로 평가할 수 있다.

연관규칙 학습기법은 *Cond*과 *Class*가 동시에 자주 등장하는 규칙을 학습한다(Agrawal and Srikant, 1994; Mirabadi and Sharifian, 2010). 이를 위해 *Cond*에 해당하는 예제 중 *Class*에 해당하는 예제의 비율을 나타내는 확신도 (Confidence)가 높은 규칙을 선호한다. 확신도가 높은 규칙 중에는 흥미로운 규칙도 있지만 *Cond*에 너무 복잡한 조건이 포함되어 일반성이 떨어지는 규칙도 있다. 따라서 학습된 규칙의 일반성을 보장하려면 규칙이 만족하는 예제의 수인 지지 값(Support)의 하한선을 정해야 한다. 이 값이 너무 높으면 조건을 만족하는 규칙을 학습하지 못할 수 있으며, 너무 낮으면 학습된 규칙의 조건이 복잡해진다. 반면 서브그룹 발견기법은 *Cond*이 주어질 때 *Class*의 분포에 큰 차이가 나는 규칙을 학습한다(Wrobel, 1997; Lavrac et al., 2004; Flach, 2012). 이를 위해 3장에 소개될 가중 상대 정확도(Weighted Relative Accuracy, *WRA*)라는 평가 함수를 이용한다(Clark and Niblett, 1989). *WRA*는 지지값의 설정 없이 확신도와 일반성을 동시에 만족하는 규칙을 찾는데 도움을 준다.

연관규칙 학습기법과 서브그룹 발견기법은 규칙을 학습하는 방식에서도 차이를 보인다. 연관규칙 학습기법은 지지값이 하한선 이상인 리터럴의 조합인 아이템(item)을 모두 찾은 후 이들을 *Cond*과 *Class*에 배분하는 식으로 규칙 후보를 생성하고 이 중 확신도가 높은 규칙을 도출한다. 따라서 사용자가 별도의 대상 속성이 결론부에 들어가도록 지정하지 않더라도 임의의 규칙들이 도출되며, 사용자가 관심 있는 대상 속성을 지정해서 규칙을 도출할 수도 있다. 반면 서브그룹 발견 기법의 경우는 대상 속성이 정해진 상태에서만 해당 속성에 대해서 *WRA*를 최대화할 수 있는 규칙을 탐색한다. 따라서 본 논문에서는 대상 속성을 지정하지 않고 규칙을 학습하는 경우와 사용자가 관심 있는 대상 속성이 있는 경우로 나누어 두 알고리즘을 적용 및 비교하였다.

규칙 학습 시 *Class*에 해당하는 대상 속성을 지정한다는 것은 그 속성의 특성에 대해 관심이 있음을 의미한다. 대상 속성은 기존의 데이터에 존재하는 속성 중 하나를 지정하거나, 사용자가 관심 있는 여러 속성들을 조합하여 만들 수 있다. 예를 들어 운전 경력과 성별의 조합에 따라 주요 사고 원인이 어떻게 다른지 알고 싶은 경우 '*Driving career* $\leq$ *7 years* $\wedge$ *Sex* = *Male*'이라는 새로운 합성 속성을 만들고 이에 대한 규칙의 학습을 시도할 수 있다. 이렇게 도출되는 규칙들은 대개 그 수가 너무 많아 사용자가 이해하기 어려우므로, 서로 중복되는 규칙을 제거하고, 여러 규칙을 일반화된 하나의 규칙으로 합치는 후처리 과정을 거친다.

본 논문의 구성은 다음과 같다. 2장에서는 사고 데이터를 이용해서 데이터 마이닝을 수행한 관련 연구에 대해서 소개하며 3장에서는 교통사고 데이터 및 데이터 마이닝을 위한 전처리 과정에 대해 소개한다. 4장에서는 데이터를 가지고 연관규칙 학습과 서브그룹 발견기법을 적용하는 방법에 대해 설명한다. 5장에서는 사고 데이터를 대상으로 데이터 마이닝 기법을 적용한 결과에 대해 논의한다. 6장에서는 논문의 내용에 대한 요약 및 결론으로 마무리한다.

2. 관련 연구

사고 데이터를 대상으로 한 데이터 마이닝 기법을 적용한 연구에는 지도 학습 기법을 적용하여 사고 피해, 사고 빈도와 다른 요소들과의 관계를 분석하는 연구가 대부분을 차지하고 있다. Chong et al.(2005)의 연구에서는 인공 신경망, 의사 결정 트리, 서포트 벡터 머신을 이용하여 교통사고 피해 예측을 수행하였고, Chang and Chen(2005)의 연구에서는 의사 결정 트리 알고리즘을 이용하여 대만 고속도로의 도로 특성과 교통사고 빈도 사이의 관계를 분석하였다. Bayam et al.(2005)의 연구에서는 고령 운전자들의 운전 특성과 사고 피해 정도의 관계를 분석하기 위해 데이터의 통계적 분석과 더불어 인공 신경망, 의사 결정 트리 알고리즘을 적용하였다. Beshah and Hill(2010)의 연구에서는 장소, 도로 특성, 조명, 날씨 정보를 기반으로 의사 결정 트리, 나이브 베이즈 모델, k 인접 이웃 알고리즘을 적용하여 사고 피해와의 관계를 분석하였다. 이와 같은 연구들은 학습 모델의 예측 성능을 극대화하는데 초점을 맞추고 있으나 학습된 모델로부터 데이터의 특성이나 속성 사이의 관계를 분석하기가 어렵다는 문제가 있다.

연관규칙 학습기법이나 서브그룹 발견기법은 이러한 연구들과는 달리 예측 성능을 극대화하는데 목적을 두지 않고 속성 사이의 관계를 잘 나타내는 규칙을 학습하는데 목적을 두고 있다. 연관규칙 학습기법은 지정된 대상 속성을 필요로 하지 않기 때문에 데이터에 대한 지식이 부족한 상태에서의 초기 연구에 활용하기 쉬우며, 사고 데이터의 경우 철도 사고 데이터를 가지고 사고에서 동시에 자주 등장하는 사고 요인들을 발견하는데 적용된 바 있다(Mirabadi and Sharifian, 2010). 한편 사용자가 데이터에 대한 지식을 어느 정도 가지고 있거나 관심 있는 대상 속성이 있는 경우 지정하는 경우 *Class*에 사용자가 지정한 대상 속성이 포함된 규칙만을 받아들이는 방식으로도 규칙을 학습할 수 있다. 서브그룹 발견기법은 연관규칙 학습기법과는 다르게 대상 속성이 정해진 상태에서만 적용 가능하지만 연관규칙 학습기법과는 다른 평가기준인 WRA를 이용하여 규칙의 일반성과 확신도를 동시에 만족하는 규칙을 찾기 때문에 고 위험군 환자 그룹 발견 (Gamberger and Lavrac, 2002), 노동자들 중 만족도에 큰 차이를 보이는 그룹 발견(Natu, and Palshikar, 2010), 소셜 미디어 마이닝을 통한 중요 인물, 정서, 집단 발견 등의 연구 (Atzmuller, 2012)과 같이 클래스 분포가 큰 차이를 보이는 집단을 찾아내는데 사용되고 있다.

3. 교통사고 데이터

본 연구에서 사용하는 교통사고 데이터에는 국내 한 대도시에서 2011년 1월부터 2014년 8월

까지 발생한 50,709건의 교통사고가 기록되어 있다. <Table 1>에서는 교통사고 데이터의 데이터 필드들을 보여주고 있다. 데이터 마이닝의 수행을 위한 전처리 과정으로써 사고 분석과 큰 관련이 없는 'Police office', 'Date', 'Geographic coordinate', 'Vehicle ID'와 같은 필드를 제거하였고, 사고 발생에 영향을 주는 것으로 알려진 'Temperature'나 'Precipitation'과 같은 기상 데이터를 사고 지점의 좌표를 기준으로 가장 가까운 관측소의 데이터를 이용하여 결합하였다. 새롭게 추가된 두 기상 속성들은 'Date'에 비해서 계절이 사고에 어떤 영향을 주는지 분석하는 데에 도움을 준다.

속성의 전처리와 더불어 데이터의 일부 필드에 대해 그룹화나 이산화 과정 또한 수행하였다. 'Day type'은 평일과 휴일로 단순화해서 구분하였고 'Time'은 주관적인 기준에 따라 <Table 2>에 나타난 5개의 구간으로 구분하였다. 'Address' 또한 대도시 내의 15개의 구 이름으로 단순화하였고, 'Place description'은 <Table 2>에 나타난 8개의 유형 중 하나의 값을 가지도록 'Place'라는 속성으로 단순화하였다. 'Vehicle type' 또한 교통법의 구분에 따라 총 10가지 유형으로 분류하였으며, 'Age'는 10년 단위로 구분하였다. 'Driving career'은 무면허 운전과 그룹별 빈도가 비슷한 5개 그룹으로 분류하였다. 'Precipitation'은 기상청의 기준에 따라 5개의 그룹으로 분류하였고, 'Temperature' 는 Brijs et al.(2008)의 구분 기준에 따라 4개 그룹으로 분류하였다.

〈Table 1〉 Data Fields of Traffic Accident Data

Data Fields	Description
Police office	Office ID that files the accident report
Date	Date / Month / Year
Day type	Day of the week
Time	Hour : Minute
Address	Address of the location of accident
Geographic coordinate	Longitude and latitude of the location of accident
Place description	Description of the place of accident
Type of accident	Vehicle-to-vehicle, Vehicle-to-human, or Others
Regulations violated	Signal violation, Lane violation, Safety duty violation, etc.
Vehicle ID	License plate number
Type of vehicle	Maker and year of the vehicles involved
Age	Age of the driver
Sex	Gender of the driver
Driving career	Year first licensed
Drunk Driving	True of False
Hit-and-run	True of False

〈Table 2〉 Attributes of Traffic Accident Data After Preprocessing

Attribute	Value
Day type	Weekday or Weekend
Time	00:00～08:00, 08:00～13:00, 13:00～17:00, 17:00～21:00, 21:00～24:00
Address	Names of 15 Regional districts
Place	8 categories (Intersection, Crosswalk, Motorway, School zone, Marketplace, Parking area, etc.)
Accident type	Vehicle-to-vehicle, Vehicle-to-human, or Others
Violation	6 types (Signal violation, Lane violation, Safety duty violation, etc.)
Vehicle type	10 types (Subcompact, Compact, Mid-sized, Full-sized, SUV, Truck, Bus, Bicycle etc.)
Age	0～19, 20～29, 30～39, 40～49, 50～59, 60+
Sex	Male or Female
Driving career	Unlicensed, Under 7 years, 7～13 years, 14～18 years, 19～23 years, Over 24 years
Drunk driving	True or False
Hit-and-run	True or False
Precipitation	None, 0～5mm, 5～20mm, 20～80mm, 80mm+
Temperature	Below zero, 0～10°C, 10～20°C, 20°C+

4. 연관규칙 학습과 서브그룹 발견기법

본 논문에서는 연관규칙 학습 알고리즘인 Apriori 알고리즘 (Agrawal and Stikant, 1994)과 서브그룹 발견 알고리즘인 MIDOS 알고리즘 (Wrobel, 1997)을 이용하여 규칙을 학습하였다. 본 장에서는 각 기법이 규칙을 학습하는 방법과 규칙을 평가하는 기준에 대해서 소개하고, 규칙 학습 시 사용자의 관심도를 반영하기 위해 대상 속성을 선정하는 방법과 학습된 규칙 집합을 단순화하여 이해하기 쉽도록 후처리 하는 과정에 대해서 설명한다.

4.1 연관규칙 학습기법

연관규칙 학습기법의 목적은 데이터 내에서 동시에 자주 등장하는 요인들의 패턴을 찾는 것이다. 이를 위해 리터럴의 교집합이 없는 두 아이템 *Cond*와 *Class*를 조합하여 'if *Cond* then *Class*' 형태의 규칙을 생성한 후 이들 중 확신도가 높은 상위 *k*개의 규칙을 도출한다. 확신도란 다음과 같이 정의할 수 있다.

$$Confidence(Cond \rightarrow Class) = Support(Cond \wedge Class) \,/\, Support(Cond)$$

여기서 *Support*(*X*)는 전체 예제 중 *X*를 만족하는 예제의 수를 나타낸다. 즉 규칙의 *Cond*를 만족하는 예제 중 *Class*도 동시에 만족하는 예제의 비중이 높을수록 확신도 값은 높아진다. 하지만 확신도가 높은 규칙을 얻기 위해 '*Day type* = *Weekend* ∧ *Place* = *Marketplace* ∧ *Vehicle type* = *Bus*' 와 같이 *Cond* 내의 리터럴 수가 많아질수록 일반성이 떨어지는 규칙을 얻게 되는 문제가 있다.

Inputs: dataset D, minimum support α, desired number of rules k, itemset size s
Outputs: rule set H
L_1 = set of itemset with size = 1 and count ≥ α
for (s = 2; $L_{s-1} \neq \emptyset$; s++)
  C_s = set of items with size = s generated by combining all non-duplicated possible pair of items in L_1 with L_{s-1}
  forall transantions $t \in D$
    C_t = subset of C_s that covers a transantion t
    forall candidates $c \in C_t$
      c.count++
  L_s = {$c \in C_s$ | c.count ≥ α}
$L = \cup_s L_s$
R = exhaustive set of '$Cond \rightarrow t$' rules generated by combining $Cond$ = {$Cond \in L$} with
t = {t ∈ L_1, $t \cap Cond = \varnothing$}
H = top k highest confidence rules in the R

〈Figure 1〉 Apriori Algorithm

<Figure 1>에서는 본 논문에서 사용한 Apriori 알고리즘을 나타내고 있다. Apriori 알고리즘에서는 도출되는 규칙들의 최소한의 일반성을 보장하기 위해 사용자가 지정한 지지값 임계치 a 이상인 아이템 집합을 도출한다. 아이템 집합의 도출은 리터럴을 하나만 가지고 있는 지지값 임계치 a 이상인 아이템 집합 L_1으로부터 L_{s-1}까지 L_1의 아이템을 추가하여 순차적으로 진행된다. 아이템 집합의 도출이 끝나면 집합 내 아이템을 조합하여 후보 규칙을 생성한 후 가장 확신도가 높은 *k*개의 규칙을 도출한다.

4.2 서브그룹 발견기법

서브그룹 발견기법은 연관규칙 학습과 마찬가지로 'if *Cond* then *Class*' 형태의 규칙을 도출하지만 연관규칙 학습기법과 다르게 *WRA*를 규칙의 평가 함수로 삼는다. *WRA*는 다음과 같이 정

의할 수 있다.

$$WRA(Cond \rightarrow Class) = p(Cond) \cdot (p(Class|Cond) - p(Class))$$

여기서 평가 식의 첫 번째 항인 $p(Cond)$은 규칙의 일반성을 나타난다. 두 번째 항은 규칙의 정확도인$p(Class|Cond)$와 $Cond$이 없을 때의 기본 정확도인 $p(Class)$의 차이로 이를 상대적 정확도라 한다. 상대적 정확도는 $Cond$으로 인해 $Class$의 데이터 분포가 얼마나 바뀌었는지를 나타낸다. WRA는 첫 번째 항인 $p(Cond)$을 일종의 가중치로 적용하여 조건이 많이 붙은 규칙은 확신도가 높더라도 낮은 $p(Cond)$을 값을 가지게 함으로써 도출된 규칙이 어느 정도의 일반성과 상대적 정확도를 동시에 갖도록 해준다.

<Figure 2>에서는 본 논문에서 사용한 서브그룹 발견 알고리즘인 MIDOS 알고리즘을 나타내고 있다. 이 알고리즘은 사용자가 지정한 대상 속성 t와 사용자가 지정한 길이 l이하의 모든

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

〈Figure 2〉 MIDOS Algorithm

$Cond$의 조합에 대한 규칙 집합인 Q를 생성하는 것으로부터 시작한다. 이 때 $Cond$이 지나치게 긴 규칙에는 사용자가 큰 관심을 갖지 않게 때문에 $Cond$의 조합을 모두 만들더라도 많은 시간이 소요되지 않는다. 규칙 생성이 끝나고 나면 while 루프에서는 Q에 있는 각 규칙들을 탐색 전략에 따라 순차적으로 평가하면서 WRA 값이 가장 높은 k개의 규칙들을 버퍼 H에 유지하게 된다.

본 논문에서는 규칙의 평가를 $Cond$의 길이가 짧은 규칙부터 순차적으로 수행하였다. 불필요한 규칙의 평가를 줄이기 위해서 Q에서 나온 규칙 r에 대해서 $WRA^*(r)$을 계산하여 r에서 $Cond$에 더 많은 리터럴이 추가된 규칙들이 평가될 필요가 있는지를 판단한다. $WRA^*(r)$는 $WRA(r)$에서 상대적 정확도 값이 최대가 되는 경우, 즉 $p(Class|Cond)$가 1이 되는 경우를 가정한 경우의 값이다. 만일 규칙 r의 $Cond$에 리터럴을 추가하여 r`을 만들 경우 $WRA(r') \leq WRA^*(r)$가 성립하는데 $r$`의 $p(Cond)$은 r보다 작으며, $p(Class|Cond)$ 또한 1보다 클 수 없기 때문이다. 따라서 어떤 규칙의 $WRA^*(r)$이 현재 버퍼 H 내의 규칙들의 최소 WRA 값 보다 작은 경우 r의 $Cond$에 리터럴을 추가하더라도 H 내의 규칙들보다 더 높은 WRA 값을 가질 수 없기 때문에 r의 $Cond$에 리터럴을 추가한 규칙들의 평가를 생략할 수 있다. WRA는 규칙 평가를 위한 일종의 휴리스틱 함수이기 때문에 MIDOS 알고리즘에 너무 적은 수의 규칙을 요청하면 흥미로운 여러 규칙들을 놓칠 수 있으며, 반대로 너무 많은 규칙을 요청하면 결과를 종합하기 어려워진다.

4.3 대상 속성의 선정

규칙 학습 시 사용자의 의도를 반영하기 위한

대상 속성은 데이터에 존재하는 속성 중 하나를 선정할 수도 있고, 관련 있는 여러 속성을 조합함으로써 만들 수도 있다. 만일 우리가 m개의 값과 n개의 값을 가지는 두 속성 F_i와 F_j를 조합한다고 가정해보자. 이 경우 두 속성을 조합한 속성은 $m \cdot n$개의 대상 값을 가질 수 있다. 각각의 대상 값에 대하여 이 값에 해당하는 경우와 그렇지 않은 두 가지로 나누는 경우 이진화된 속성 F_t를 얻을 수 있다. 이제 데이터에서 F_i와 F_j를 제외하고 F_t를 추가한 후 대상 속성으로 지정하면 규칙 학습 기법을 적용하여 규칙을 도출할 수 있다. 이러한 대상 속성을 만드는 경우의 수는 조합된 속성의 값을 어떻게 분할하느냐의 경우와 수와 동일하다. 예를 들어 조합된 속성의 값이 a, b, c의 3가지인 경우 a, b, c를 별도로 보는 경우와 a와 b, c로 나누는 경우, b와 a, c로 나누는 경우, c와 a, b로 나누는 경우의 총 4가지 조합이 가능하다. 대상 속성을 선정한 후 규칙을 도출하고 나면 규칙 집합 내의 규칙 수가 많거나 비슷한 규칙이 많이 있을 수 있다. 따라서 후처리 과정을 통해 도출된 규칙 집합을 단순화하여 이해하기 쉽게 만드는 과정이 필요하다.

4.4 후처리 과정

후처리 과정의 첫 단계에서는 충분히 검증되지 않았거나 흥미롭지 않은 규칙들을 제거한다. 충분히 검증된 규칙만을 유지내기 위해서 규칙이 만족하는 예제의 수인 지지값이 사용자가 지정한 임계치 β보다 낮은 모든 규칙을 제거한다. 흥미도 보장을 위해서는 $p(Class) / p(Class|Cond)$로 정의되는 lift 값이 임계치 γ보다 낮은 모든 규칙을 제거한다. 다음 단계에서는 불필요(redundant)한 규칙들을 제거한다. 어떤 규칙 A와 다른 규칙 B가 있을 때 $Cond_A \cap Cond_B = Cond_A$이고 $p(Class|Cond_B) / p(Class|Cond_A)$ 임계치 δ보다 낮으면 규칙 B는 불필요한 규칙이라고 볼 수 있다. 그 이유는 규칙 B는 A에 비해 $Cond$에 추가적인 리터럴을 갖고 있음에도 불구하고 정확도의 증가량이 기대치만큼 높지 않기 때문이다. 마지막 단계로 $Cond$에 동일한 속성들을 가진 규칙들을 속성 별로 비접속 형태(disjunctive form)의 값을 가지도록 하나의 규칙으로 병합한다. 예를 들어 $F_2 = a \wedge F_3 = b$ 와 $F_2 = a \wedge F_3 = c$ 는 $F_2 = a \wedge F_3 = \{b, c\}$로 병합할 수 있다. 모든 후처리 과정이 끝나고 나면 남은 모든 규칙들은 사용자의 편의를 위해 WRA 순으로 정렬하여 표시한다.

5. 실험 결과

사고데이터 마이닝을 위한 서브그룹 발견기법과 연관규칙 학습기법의 비교를 위해 몇 가지 실험을 수행하였다. 첫 번째 실험에서는 사고에서 자주 등장하는 패턴이 무엇인지를 관찰하기 위해 대상 속성을 지정하지 않고 연관규칙 학습기법을 적용하여 규칙을 도출하였다. 두 번째 실험에서는 대상 속성이 지정된 경우 두 규칙 학습 알고리즘이 어떤 차이를 보이는지 확인한기 위해 ‘Driving Career’를 대상 속성으로 서브그룹 발견기법과 연관규칙 학습기법을 적용하였다. 마지막 실험에서는 대상 속성의 결합이 학습 결과에 어떤 변화를 일으키는지 확인하기 위해 ‘Driving Career’와 ‘Sex’를 결합하여 서브그룹 발견기법과 연관규칙 학습기법을 적용하여 발견한 규칙을 분석하였다.

5.1 연관규칙 학습기법을 이용한 규칙 학습

사고 데이터에 자주 등장하는 패턴을 발견하기 위해 연관규칙 학습기법을 적용하여 규칙을 학습하였다. item의 최소 지지값 a는 학습할 규칙의 일반성을 보장하기 위해 전체 데이터의 약 5%에 해당하는 2,500으로 지정하였다. item의 조합으로 생성 가능한 규칙들 중 흥미로운 규칙만을 얻기 위해 확신도가 높은 500개의 규칙을 선별한 후 Lift 값이 1.3 이하인 규칙을 제거한 결과 총 98개의 규칙을 얻을 수 있었다. 하지만 모든 규칙이 'Age'와 'Driving career'에 대한 연관관계를 포함한 규칙이었다. 운전경력과 연령대가 높은 연관성을 가지는 것은 매우 당연한 사실이다. 따라서 'Age' 속성을 제외하고 다시 규칙 학습을 수행하였다. 그 결과 총 52개의 규칙을 얻을 수 있었다.

<Table 3>은 도출된 52개의 규칙 중 대표적인 일부 규칙들을 선별하여 표시한 것이다. 도출된 규칙들 중에는 교차로에서는 차:차 사고가 많이 발생하는 것이나 화물차 사고는 평일에 남성 운전자가 많이 일으키는 것과 같이 이미 잘 알려져 있거나 당연한 사실들이 많이 포함되어 있다. 반면에 운전 경력과 사고의 관계를 나타낸 흥미로운 규칙들도 있었다. <Table 3>의 규칙들을 보면 24년 이상의 운전 경력은 남성 운전자의 중형차 사고나 남성 운전자의 오전 시간대 사고와 연관성이 높음을 알 수 있다.

위 분석 결과를 토대로 운전 경력이 사고 발생에 어떤 영향을 주는지 확인하기 위해 'Driving career'를 대상 속성으로 하여 규칙을 학습함으로써 추가적인 분석을 수행하기로 결정하였다. 이와 같이 연관규칙 학습기법은 대상 속성을 지정하지 않아도 흥미로운 패턴을 발견하는데 도움이 되며 심화된 분석을 위한 대상 속성의 선정에도 도움이 되는 것을 알 수 있다.

〈Table 3〉 Rules Discovered by Association Rule Learning

Condition	Class	Support	Lift
Driving career = Over 24 years	**Vehicle type = mid-sized, Sex = Male**	4,866	1.4
Place = Intersection	Accident type = Vehicle-to-vehicle	5,819	1.38
Vehicle type = Truck	Day type = Weekday, Sex = Male	4,227	1.38
Vehicle type = Mid-sized, Sex = Male	Time = 00:00 ~ 08:00	4,402	1.38
Time = 08:00 ~ 13:00, Sex = Male	**Driving Career = Over 24 years**	2,599	1.36
Day type = Weekend, Sex = Male	Time = 00:00 ~ 08:00	3,311	1.32

5.2 'Driving career' 대상의 규칙 학습

사고 데이터의 운전자 경력별 분포는 총 50,709건의 사고 중 무면허 3,719건 (7.3%), 7년 미만 9,309건 (18.4%), 7 ~ 13년 10,354건 (20.4%), 14 ~ 18년 6,664건 (13.1%), 19년 ~ 23년 9,039건 (17.8%), 24년 이상 11,624건 (22.9%)으로 구성되어 있다. 이를 통해 그룹별 분포는 비교적 고르며, 무면허 운전 사고도 상당한 비중을 차지함을 알 수 있다. 각 운전경력 그룹의 사고 특성을 알아보기 위해 6개의 운전 경력별 그룹을 대상 속성으로 하여 연관규칙 학습기법과 서브그룹 발견기법을 적용하였다.

연관규칙 학습기법의 경우 아이템의 최소 지지값 a에 대하여 전체 데이터 수 대비 5%, 1%, 0.5%, 0.1%의 비율 이상이 되도록 2,500, 500,

250, 50의 네 가지 조합을 시도하였고, 6개의 운전자 경력 그룹을 대상 속성으로 하여 각 100개씩 600개의 규칙을 도출하였다. 도출된 규칙들 중 일반성이 떨어지거나 흥미도가 낮은 규칙을 제외하기 위해 지지값이 100 이하이거나 lift가 1.5 이하인 규칙들을 제거하고 규칙의 후처리 과정을 적용하였다.

실험 결과 a를 2,500이나 500으로 한 경우에는 규칙을 발견할 수 없었다. a를 250으로 한 경우 확신도가 높은 600개의 규칙 중 조건을 만족하는 218개의 후보 규칙을 얻을 수 있었다. 후보 규칙에 후처리 과정을 적용한 결과 redundant 한 규칙 제거로 25개의 규칙을, 규칙 병합을 통해 최종 23개의 서브그룹을 얻을 수 있었다. 이를 통해 후처리 과정이 서로 비슷한 규칙들을 줄이는 데 효과적임을 알 수 있다. 한편 a를 50으로 한 경우 확신도가 높은 600개의 규칙 중 후처리를 거쳐 최종 57개의 규칙을 얻을 수 있었으나 a를 250으로 한 경우에 비해 도출된 규칙들의 길이가 길어서 분석이 어려워지는 문제가 있었다. 이는 a 값이 낮아질수록 좀 더 다양한 조합의 규칙을 생성할 수 있지만 그만큼 길이가 길면서 일반성이 떨어지는 규칙들이 높은 확신도 때문에 상위권에 들어가는 경우가 많기 때문이다.

서브그룹 발견기법의 경우 각 대상 속성별로 100개씩 총 600개의 규칙을 *WRA*가 높은 순으로 도출하였다. 연관규칙 학습기법과 마찬가지로 도출된 규칙 중 지지값이 100 이하이거나 lift가 1.5 이하인 규칙들을 제거하고 후처리 과정을 적용하였다. 실험 결과 조건을 만족하는 184개의 규칙 중 후처리 과정을 통해 최종 21개의 규칙을 얻을 수 있었다. 두 기법이 도출한 서브그룹의 비교를 위해 공통으로 찾아낸 서브그룹들 중 의미 있는 일부 서브그룹을 <Table 4>에, 각 기법에서 고유하게 찾아낸 서브그룹을 각각 <Table 5, 6>에 나타내었다.

〈Table 4〉 Subgroups for Different Driving Career (Both Association Rule Learning and Subgroup Discovery)

Subgroup	Support	Lift	*WRA*
Driving career = Unlicensed			
Vehicle type = Bicycle	415	1.66	3.2.E-03
Driving career = Under 7 years			
Vehicle type = {Subcompact, bicycle}	2,159	1.85	2.0.E-02
Drunk driving = true	1,425	1.73	1.2.E-02
Accident type = Vehicle-to-vehicle, Vehicle type = Compact	576	1.54	4.0.E-03
Driving career = 7 ~ 13 years			
Vehicle type = Compact	881	1.50	5.8.E-03
Sex = Female, Day type = Weekend	600	1.51	4.0.E-03
Sex = Male, Time = 00:00 ~ 08:00, Drunk driving = true	593	1.51	3.9.E-03
Time = 00:00 ~ 08:00, Accident type = Vehicle-to-vehicle, Drunk driving = true	537	1.50	3.5.E-03
Driving career = 14 ~ 18 years			
Sex = Female	1664	1.56	1.2.E-02
Driving career = 19 ~ 23 years			
Driving career = Over 24 years			
Vehicle type = {Bus, Special purpose car}	1,568	2.40	1.8.E-02
Sex = Male, Accident type = Vehicle-to-human, Vehicle type = Mid-sized	1,562	1.60	1.2.E-02
Sex = Male, Time = 08:00 ~ 12:00, Vehicle type = Mid-sized	957	1.58	7.0.E-03
Time = 00:00 ~ 08:00, Accident type = Vehicle-to-human, Vehicle type = Mid-sized	499	1.64	3.9.E-03

〈Table 5〉 Subgroups for Different Driving Career (Subgroup Discovery Only)

Subgroup	Support	Lift	*WRA*
Driving career = 19 ~ 23 years			
Sex = Female, Vehicle type = Full-sized	221	1.53	1.5.E-03

〈Table 6〉 Subgroups for Different Driving Career (Association Rule Learning Only)

Subgroup	Support	Lift	*WRA*
Driving career = Under 7 years			
Day type = Weekend, Vehicle type = Compact	269	1.68	2.1E-03
Time = 00:00 ~ 08:00, Vehicle type = SUV	275	1.51	1.8E-03
Driving career = Over 24 years			
Sex = Male, Time = 13:00 ~ 17:00, Day type = Weekday, Vehicle type = Mid-sized	515	1.59	3.8.E-03

두 기법은 전반적으로 비슷한 규칙들을 찾아내지만 연관규칙 학습기법에서는 운전경력이 19 ~ 23년인 경우 조건을 만족하는 규칙을 학습하지 못한 것을 볼 수 있다. 이는 서브그룹 발견기법에서 찾아낸 '*Sex = Female, Vehicle type = Full-sized*'라는 서브그룹은 조건을 만족하는 예제의 수가 221개이기 때문에 연관규칙 학습기법에서 a를 250 이상인 아이템만을 가지고 규칙을 만드는 경우 해당 조건은 포함되지 못하기 때문이다. 이를 통해 연관규칙 학습기법은 a 값이 학습 결과에 큰 영향을 미친다는 것을 확인할 수 있다.

두 기법이 찾아낸 서브그룹들을 살펴보면 운전경력에 따라서 사고가 많이 발생하는 조건이 달라짐을 확인할 수 있다. 운전경력이 13년 미만인 운전자들은 소형차나 이륜차의 사고가 많고 음주운전 사고도 많이 일으킨다. 이는 운전경력이 적은 사람들 중 젊은 사람이 많기 때문인 것으로 추정된다. 운전경력이 14 ~ 23년 사이의 운전자 사고는 여성 운전자가 많았으며, 운전경력이 24년 경우 긴 운전경력을 필요로 하는 특수차량의 사고나 아침 시간대의 사고나 중형차 운전자의 차:사람 사고가 많아지는 것을 확인할 수 있다.

5.3 'Driving career'와 'Sex'의 조합

운전 경력을 대상 속성으로 한 실험에서 경력별로 사고를 많이 일으키는 성별이 달라진다는 사실을 확인할 수 있었다. 따라서 성별과 운전경력에 대한 심층적인 분석을 위해서 두 속성을 결합하여 대상 속성을 만든 후 두 가지 규칙 학습기법을 적용하여 비교하였다.

연관규칙 학습기법의 경우 아이템의 최소 지지값 a에 대하여 전체 데이터 수 대비 0.05, 0.01, 0.005, 0.001의 비율 이상이 되도록 2,500, 500, 250, 50의 네 가지 조합을 시도하였고, 6개의 운전자 경력 그룹과 2개의 운전자 성별의 조합을 대상 속성으로 하여 각 100개씩 총 1,200개의 규칙을 도출하였다. 도출된 규칙들 중 일반성이 떨어지거나 흥미도가 낮은 규칙을 제외하기 위해 지지값이 100 이하이거나 lift가 1.5 이하인 규칙들을 제거하고 규칙의 후처리 과정을 적용하였다. 실험 결과 a를 2,500이나 500으로 한 경우에는 규칙을 발견할 수 없었다. a를 250으로 한 경우 확신도가 높은 1,200개의 규칙 중 조건을 만족하는 111개의 후보 규칙을 얻을 수 있었으며 후처리 과정을 통해 최종 19개의 서브그룹을 얻을 수 있었으나 서브그룹을 도출하지 못하는 조합이 많아 채택하지 않았다. a를 50으로 한 경우 확신도가 높은 1,200개의 규칙 중 후처리를

거쳐 최종적으로 38개의 서브그룹을 얻을 수 있었다.

서브그룹 발견기법의 경우 각 대상 속성별로 100개씩 총 1,200개의 규칙을 *WRA*가 높은 순으로 도출하였다. 연관규칙 학습기법과 마찬가지로 도출된 규칙 중 지지값이 100 이하이거나 lift가 1.5 이하인 규칙들을 제거하고 후처리 과정을 적용하였다. 실험 결과 조건을 만족하는 238개의 규칙을 도출할 수 있었으며 후처리 과정을 통해 최종 40개의 서브그룹을 얻을 수 있었다.

두 기법이 도출한 서브그룹의 비교를 위해 공통으로 찾아낸 서브그룹들 중 의미 있는 일부 서브그룹을 <Table 7>에, 각 기법에서 고유하게 찾아낸 서브그룹을 각각 <Table 8, 9>에 나타내었다. 두 기법은 앞서 실험과 마찬가지로 전반적으로 비슷한 서브그룹들을 찾아내지만 각 기법이 고유하게 찾아낸 서브그룹의 품질에 차이를 보임을 알 수 있다. 예를 들어 운전경력 7년 미만의 남성 운전자 사고에 대한 서브그룹을 보면 서브그룹 발견기법은 '*Drunk drivng = true*', '*Hit-and-run = true*'와 같이 단일 리터럴로 구성된 서브그룹을 찾아낸 반면 연관규칙 학습기법은 '*Time = 00:00 ~ 08:00, Drunk driving = true, Temperature = 10 ~ 20℃*'와 같이 음주 운전과 몇 개의 추가 리터럴로 구성된 서브그룹을 찾아내었다. 이는 확신도가 높은 단일 리터럴을 포함하면서 몇 가지 리터럴이 추가된 서브그룹들은 높은 확신도를 받기 쉬운 경향이 있기 때문이다. 반면 서브그룹 발견기법은 WRA의 적용을 통해서 일반성을 떨어트리면서까지 확신도를 높이는 서브그룹에 대한 패널티를 주기 때문에 이러한 현상을 막을 수 있었다.

운전 경력과 성별을 조합한 대상 속성에 대해서 도출한 서브그룹들을 보면 운전 경력만을 대

〈Table 7〉 Subgroups for Different Driving Career and Sexes (Both Association Rule Learning and Subgroup Discovery)

Subgroup	Support	Lift	*WRA*
Driving career = Under 7 years, Sex = Female			
Vehicle type = {Subcompact, Compact}	638	3.56	9.0.E-03
Driving career = 7 ~ 13 years, Sex = Female			
Vehicle type = {Subcompact, Compact}	774	2.94	1.0.E-02
Time = 08:00 ~ 12:00, Vehicle type = Mid-sized	237	1.52	1.6.E-03
Driving career = 14 ~ 18 years, Sex = Female			
Vehicle type = {Subcompact, Compact}	501	2.62	6.1.E-03
Time = {08:00 ~ 12:00, 13:00 ~ 17:00}, Vehicle type = Mid-sized	343	1.62	2.6.E-03
Driving career = 19 ~ 23 years, Sex = Female			
Vehicle type = {SUV, Subcompact, Compact}	672	1.82	6.0.E-03
Time = {08:00 ~ 12:00, 13:00 ~ 17:00}, Vehicle type = Mid-sized	376	1.67	3.0.E-03
Driving career = Over 24 years, Sex = Female			
Vehicle type = Full-sized	150	2.70	1.9.E-03
Time = 13:00 ~ 17:00, Day type = Weekday	140	1.55	9.8.E-04
Driving career = Unlicensed, Sex = Male			
Vehicle type = bicycle	375	3.23	5.1.E-03
Hit-and-run = True	267	4.04	4.0.E-03
Drunk driving = true	331	2.18	3.5.E-03
Driving career = Under 7 years, Sex = Male			
Vehicle type = bicycle	1254	2.43	1.5.E-02
Driving career = 14~18 years, Sex = Male			
Accident type = Vehicle-to-human, Vehicle type = Truck	234	1.51	1.6.E-03
Driving career = 19 ~ 23 years, Sex = Male			
Vehicle type = {Van, Truck}	1681	1.52	1.1.E-02
Day type = Weekday, Vehicle type = Bus	308	1.52	2.1.E-03
Driving career = Over 24 years, Sex = Male			
Vehicle type = {Bus, Special purpose car}	1561	2.54	1.9.E-02
Time = 00:00 ~ 08:00, Accident type = Vehicle-to-human, Vehicle type = Mid-sized	485	1.69	3.9.E-03

〈Table 8〉 Subgroups for Different Driving Career and Sexes (Association Rule Learning Only)

Condition	Support	Lift	*WRA*
Driving career = Under 7 years, Sex = Male			
Time = 00:00 ~ 08:00, Drunk driving = true, Temperature = 10 ~ 20℃	268	2.41	3.1.E-03
Driving career = 7 ~ 13 years, Sex = Male			
Drunk driving = true, Vehicle type = {SUV, Mid-sized}	695	1.82	6.2.E-03
Drunk driving = true, Time = 00:00 ~ 08:00	593	1.74	5.0.E-03
Hit-and-run = True, Vehicle type = Mid-sized, Accident type = Vehicle-to-vehicle	153	1.71	1.3.E-03
Driving career = Over 24 years, Sex = Female			
Vehicle type = Bus	1157	2.84	1.5.E-02
Time = 08:00 ~ 12:00, Vehicle type = Special purpose car	159	2.13	1.7.E-03

〈Table 9〉 Subgroups for Different Driving Career and Sexes (Subgroup Discovery Only)

Condition	Support	Lift	*WRA*
Driving career = Under 7 years, Sex = Male			
Drunk driving = true	1,284	1.90	1.2.E-03
Hit-and-run = true	485	1.65	3.8.E-03
Driving career = 7 ~ 13 years, Sex = Male			
Drunk driving = true	1,138	1.60	8.4.E-03
Hit-and-run = True, Vehicle type = Mid-sized	177	1.58	1.3.E-03
Time =21:00 ~ 24:00, Vehicle type = SUV	168	1.50	1.1.E-03
Driving career = Over 24 years, Sex = Female			
Vehicle type = {Bus, Special purpose car}	1561	2.54	1.9.E-02
Time = 00:00 ~ 08:00, Accident type = Vehicle-to-human, Vehicle type = Mid-sized	485	1.69	3.9.E-03

상 속성으로 한 경우보다 추가적인 정보를 얻을 수 있다. 예를 들어 <Table 7, 9>의 13년 미만의 남성 운전자 사고를 보면 공통적으로 음주운전과 뺑소니 사고를 많이 일으키는 것을 볼 수 있다. 이는 운전 경력만을 대상으로 했을 때 찾아낸 13년 미만의 운전 경력을 가진 운전자가 음주운전을 많이 한다는 것보다 상세화된 정보이다. 한편 여성 운전자의 경우에는 경차나 소형차 운전자 사고가 많으며 아침부터 저녁 이전 시간까지 주로 사고가 발생하는 것을 볼 수 있다.

6. 결론

본 논문에서는 여러 속성으로 이루어진 사고 데이터를 이용하여 규칙 기반 마이닝 기법인 연관규칙 학습기법과 서브그룹 발견기법을 적용하였다. 연관규칙 학습기법은 대상 속성이 지정되지 않은 경우에도 속성 사이의 관계를 발견할 수 있어 데이터를 이용한 초기 분석에 유용함을 확인하였다. 한편 사용자의 관심사를 마이닝에 반영하고자 하는 경우 관련 있는 몇 가지 속성을 조합하여 새로운 대상 속성을 만든 후 대상 속성을 target class로 하는 규칙을 학습하였다. 대상 속성이 정해진 경우 연관규칙 학습기법은 최소 지지값에 따라 도출되는 규칙에 큰 차이를 보였으며, 대상 속성의 조합의 증가하면서 대상 속성별 데이터 분포의 차이가 점점 커지기 때문에 최적의 지지값을 찾는데 어려움을 겪었다. 반면 서브그룹 발견기법은 *WRA*의 이용한 규칙의 평가로 규칙의 일반성과 확신도를 동시에 고려한 규칙들을 학습할 수 있었다.

참고문헌(References)

Agrawal, R. and R. Srikant, "Fast algorithms for mining association rules," *Proceedings of the 20th Very Large Data Bases Conference,* (1994), 487～499.

Atzmuller, M., "Mining Social Media: Key Players, Sentiments, and Communities," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol.2, No.5 (2012), 411～419.

Bayam, E., J. Liebowitz., and W. Agresti, "Older Drivers and Accidents: A Meta Analysis and Data Mining Application on Traffic Accident Data," *Expert Systems with Applications*, Vol.29, No.3(2005), 598～629.

Beshah, T. and S. Hill, "Mining Road Traffic Accident Data to Improve Safety: Role of Road Related Factors on Accident Severity in Ethiopia," *Proceedings of the 2010 AAAI Spring Symposium Series,* (2010), 14～19.

Brijs, T., D. Karlis., and G. Wets, "Studying the Effect of Weather Conditions on Daily Crash Counts using a Discrete Time-series Model," *Accident Analysis & Prevention*, Vol.40, No.3(2008), 1180～1190.

Chang, L. and W. Chen, "Data Mining of Tree-based Models to Analyze Freeway Accident Frequency," *Journal of Safety Reserch*, Vol.36, No.4(2005), 365～375.

Chong, M., A. Abraham, and M. Paprzycki, "Traffic Accident Analysis Using Machine Learning Paradigms," Informatica, Vol 29(2005), 89～98

Clark, P., and T. Niblett, "The CN2 Induction Algorithm," *Machine Learning*, Vol.3, No.4 (1989), 261～283.

Depaire, B., G. Wets, and K. Vanhoof, "Traffic Accident Segmentation by Means of Latent Class Clustering," *Accident Analysis & Prevention*, Vol.40, No.4(2008), 1257～1266.

Flach, P, *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*, Cambridge University Press, New York, 2012.

Gamberger, D. and N. Lavrac, "Expert-guided Subgroup Discovery: Methodology and Application," *Journal of Artificial Intelligence Research*, Vol.17, No.1(2002), 501～527.

Lavrac, N., B. Kavsek., P. Flach., and L. Todorovski, "Subgroup Discovery with CN2-SD," Journal of Machine Learning Research, Vol.5(2004), 153～188.

Mirabadi, A. and S. Wets, "Application of Association Rules in Iranian Railways (RAI) Accident Data Analysis," *Safety Science*, Vol.48, No.10(2010), 1427～1435.

Natu, M. and G. K. Palshikar, "Interesting subset discovery and its application on service processes," *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops,* (2010), 1061～1068.

Witten, I. H., et al., *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)*. Morgan Kaufmann, Boston, 2011

Wrobel, S., "An algorithm for multi-relational discovery of subgroups," *Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery*, (1997), 78～87.

Abstract

Comparison of Association Rule Learning and Subgroup Discovery for Mining Traffic Accident Data

Jeongmin Kim* · Kwang Ryel Ryu**

Traffic accident is one of the major cause of death worldwide for the last several decades. According to the statistics of world health organization, approximately 1.24 million deaths occurred on the world's roads in 2010. In order to reduce future traffic accident, multipronged approaches have been adopted including traffic regulations, injury-reducing technologies, driving training program and so on. Records on traffic accidents are generated and maintained for this purpose. To make these records meaningful and effective, it is necessary to analyze relationship between traffic accident and related factors including vehicle design, road design, weather, driver behavior etc. Insight derived from these analysis can be used for accident prevention approaches. Traffic accident data mining is an activity to find useful knowledges about such relationship that is not well-known and user may interested in it.

Many studies about mining accident data have been reported over the past two decades. Most of studies mainly focused on predict risk of accident using accident related factors. Supervised learning methods like decision tree, logistic regression, k-nearest neighbor, neural network are used for these prediction. However, derived prediction model from these algorithms are too complex to understand for human itself because the main purpose of these algorithms are prediction, not explanation of the data. Some of studies use unsupervised clustering algorithm to dividing the data into several groups, but derived group itself is still not easy to understand for human, so it is necessary to do some additional analytic works.

Rule based learning methods are adequate when we want to derive comprehensive form of knowledge about the target domain. It derives a set of if-then rules that represent relationship between the target feature with other features. Rules are fairly easy for human to understand its meaning therefore it can help provide insight and comprehensible results for human. Association rule learning methods and

* Department of Electrical and Computer Engineering, Pusan National University
** Corresponding Author: Kwang Ryel Ryu
Department of Electrical and Computer Engineering, Pusan National University
2 Busandaehak-ro, 63beon-gil, Geumjeong-gu, Busan 609-735, Korea
Tel: +82-51-510-2453, Fax: +82-51-517-2431, E-mail: krryu@pusan.ac.kr

subgroup discovery methods are representing rule based learning methods for descriptive task. These two algorithms have been used in a wide range of area from transaction analysis, accident data analysis, detection of statistically significant patient risk groups, discovering key person in social communities and so on.

We use both the association rule learning method and the subgroup discovery method to discover useful patterns from a traffic accident dataset consisting of many features including profile of driver, location of accident, types of accident, information of vehicle, violation of regulation and so on. The association rule learning method, which is one of the unsupervised learning methods, searches for frequent item sets from the data and translates them into rules. In contrast, the subgroup discovery method is a kind of supervised learning method that discovers rules of user specified concepts satisfying certain degree of generality and unusualness. Depending on what aspect of the data we are focusing our attention to, we may combine different multiple relevant features of interest to make a synthetic target feature, and give it to the rule learning algorithms. After a set of rules is derived, some postprocessing steps are taken to make the ruleset more compact and easier to understand by removing some uninteresting or redundant rules.

We conducted a set of experiments of mining our traffic accident data in both unsupervised mode and supervised mode for comparison of these rule based learning algorithms. Experiments with the traffic accident data reveals that the association rule learning, in its pure unsupervised mode, can discover some hidden relationship among the features. Under supervised learning setting with combinatorial target feature, however, the subgroup discovery method finds good rules much more easily than the association rule learning method that requires a lot of efforts to tune the parameters.

Key Words : data mining, association rule learning, subgroup discovery, traffic accident data, rule learning

Received : September 27, 2015 Revised : December 7, 2015 Accepted : December 8, 2015
Corresponding Author : Kwang Ryel Ryu

저 자 소 개

김 정 민

2010년에 부산대학교 전자전기컴퓨터공학부에서 학사 학위를 취득하였고, 현재 부산대학교 전기컴퓨터공학과에서 박사 과정에 재학 중이다. 주요 연구 분야는 기계 학습, 진화형 알고리즘, 데이터 마이닝, 확률적 추론 모델이다.

류 광 렬

1981년에 서울대학교 전자공학과에서 학사 및 석사 학위를 취득하였고, 1992년에 미국 미시건 대학교의 컴퓨터공학과에서 박사 학위를 취득하였으며, 1993년부터 부산대학교 전기컴퓨터공학과에서 교수로 재직 중이다. 주요 연구 분야는 기계 학습, 진화형 알고리즘, 데이터 마이닝, 확률적 추론 모델, 지능형 스케줄링 등이다.