KIPS Tr. Software and Data Eng.
Vol.3, No.1 pp.37~42 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2014.3.1.37

A New Importance Measure of Association Rules Using Information Theory

Chang-Hwan Lee† · Joohyun Bae††

ABSTRACT

The abstract should concisely state what was done, how it was done, principal results, and their significance. It should be less than 300 words for all forms of publication. The abstract should be written as one paragraph and should not contain tabular material or numbered references. At the end of abstract, keywords should be given in 3 to 5 words or phrases.

Keywords : Association, Classification, Rule Importance, Hellinger Divergence

정보이론에 기반한 연관 규칙들의 새로운 중요도 측정 방법

이 창 환† · 배 주 현††

요 약

연관 규칙들을 이용한 분류학습은 최근 활발히 연구되는 분야의 하나이다. 이러한 연관 규칙을 이용한 분류에는 연관 규칙들에 대한 수치적 중요도를 계산하는 것이 중요하다. 본 논문에서는 정보 이론을 사용한 H measure 라는 새로운 규칙 중요도 기법을 제안한다. 구체적으로 Hellinger 변량을 이용하여 연관규칙의 중요도를 계산한다. 제안된 H measure 의 다양한 특성들을 분석하였으며 또한 이러한 H measure를 이용한 분류학습의 성능을 다른 규칙 measure를 이용한 분류학습의 성능과 비교하였다.

키워드 : 연관 규칙, 분류 학습, 규칙 중요도, 헬링거 변량

1. 서 론

연관규칙들은 “조건→결과” 형태로 표현되며 여기에서 조건과 결과는 관측값 들의 결합으로 나타난다. 이러한 연관규칙은 다른 학습 방법에 비하여 학습 결과의 이해가 쉬우며 사용자들에게 많은 정보를 준다는 이점을 가진다.

하지만 연관규칙을 추출해내는 대부분 알고리즘들은 흔히 중복성을 가지는 방대한 양의 규칙 들을 생성한다. 그러나 데이터로부터 생성된 많은 연관규칙들 중 다수는 실제로 중요하지 않으며, 이와 같이 방대한 유도된 규칙 들 중에서 중요한 규칙들의 조합을 선택하는 것은 쉽지 않은 일이다.

그러므로 연관법칙의 생성에서 각 법칙에 대하여 수치적으로 표현이 가능한 규칙의 중요도를 계산하는 것은 중요한 의미를 가진다. 이는 사용자가 발견된 규칙을 정렬하거나 분류학습을 진행할 때 각 클래스 마다 점수를 부여하는 것을 돕는다. 하지만 많은 사용자의 요구들을 충족시키는 규칙의 중요도를 계산하는 것은 쉬운 일이 아니며, 많은 연구들이 이 분야에서 시행되어 왔다.

규칙 중요도 들에는 크게 주관적인 중요도와 객관적인 중요도의 두 가지가 있다. 주관적 중요도는 사용자의 도메인(domain)지식을 고려하는 것이며 [1] 객관적인 중요도는 데이터를 사용하여 중요도를 자동으로 계산한다 [2] [3] [4] [5]. 본 연구에서는 정보이론(information-theoretic)을 이용하여 연관법칙의 새로운 객관적인 중요도를 제안하고자 한다. 정보이론을 이용한 중요도의 측정은 연관법칙에서 제공하는 정보의 양으로서의 해석이 가능하다. 즉 규칙으로서의 선행조건이 후행결과에 대해 많은 정보를 제공할 때 관계는 흥미를 가진다. 규칙 중요도를 측정하기 위해 사용된 정보이론 방법으로는 Shannon 조건부 엔트로피[6], 평균 상호정보(mutual information) [7], Theil 불확정도 계수 [4], J-measure [8], 지니 계수 [2] 등이 있다.

Shannon 엔트로피는 선행조건(antecedent)이 사실이라고 가정하였을 때 후행결과(consequent)의 정보량의 평균을 측

※ 본 연구는 한국과학재단 일반연구비 지원(2011-0023296)으로 이루어졌음.
† 종신회원 : 동국대학교 정보통신학과 교수
†† 정 회 원 : 동국대학교 정보통신학과 강사
논문접수 : 2013년 8월 9일
수 정 일 : 1차 2013년 10월 10일
심사완료 : 2013년 11월 5일
* Corresponding Author : Chang-Hwan Lee(chlee@dgu.ac.kr)

정한다 [6]. Theil 계수는 선행조건으로 인한 후행결과의 엔트로피 감소율을 측정한다 [4]. J-measure는 선행조건의 사실여부에 따른 평균 상호정보의 일부분이다 [8]. 마지막으로 지니계수는 2차 엔트로피의 감소를 측정한다 [2]. 본 연구에서는 Hellinger 함수를 이용한 새로운 중요도를 제안하고 이의 특성을 분석한다.

2. 관련 연구

연관법칙의 중요도 분야에는 다양한 연구가 진행되어 있다. 먼저 Tan et al. [4] 은 다수의 법칙 중요도에 관한 광범위한 검토를 하였다. 그들은 20가지의 중요도에 대하여 내부 자료구조, 연산, 크기에 따른 효과를 검사하였다. Tan 등은 연관규칙의 가능한 도메인 유형이 주어진 상태에서 measure 들의 주요 특징과 적용성에 대해 몇 가지 결론을 내렸다.

베이시안 네트워크를 사용하여 자주 나타나는 항목들의 흥미도를 평가한 방식은 Jarosezvicz 와 Simovici [9] 에 의해 제안되었다. 이는 베이시안 네트워크를 사용하여 도메인 지식을 제공하는 방식으로 중요도의 기준은 자료에서 파생된 support 와 베이시안 네트워크에서 계산된 수치 사이의 절대적 차이에 따라 정의된다.

앞 절에서 언급된 바와 같이, 연관규칙은 분류학습에도 적용 될 수 있으며 이 경우에 법칙의 중요도는 성능에 많은 영향을 미친다. 따라서 연관 규칙 마이닝과 분류를 통합하는데 많은 연구들이 제안되어 있다.

Dong et al. [10] 은 CAEP(Classification by Aggregating Emerging Patterns) 모델을 제안하였는데, 이는 모든 패턴을 찾은 후 점수를 차별화하여 모두 더하고 주어진 데이터에 가장 적합한 클래스를 정하는 것이다. 점수는 패턴들의 중요도를 뜻하며, 이는 support 들과 반대 클래스들의 큰 차이에서 파생된다.

Li et al. [11] 은 CMAR(Classification based on Multiple Association Rules) 방법을 제안했는데, 이는 효과적인 FPtree 구조를 이용하여 규칙을 생성하고 평가한다. CMAR 은 인스턴스에 부합하는 모든 규칙을 추출하고 가중된 카이제곱 방법을 이용하여 목적 클래스의 값을 계산한다.

3. H measure

연관규칙은 $b \rightarrow a$ 형식으로 지식을 표현한다. 본 논문은 연관 규칙을 이용한 분류학습을 다루므로 연관 규칙의 오른쪽 부분은 목적속성 값을 가정하고 좌변은 다수의 속성값으로 가정한다.

본 연구에서 규칙 중요도의 기본 개념은 연관규칙 좌변의 특정한 값 집합들이 우측 목적속성의 확률 분포에 영향을 끼친다는 가정에서 출발한다. 목적 속성은 클래스 빈도를 나타내는 이전 확률(prior distribution)을 구성한다. 하지만 목적 속성은 다른 속성들의 값 등 특정 조건 하에서 그 값들의 확률분포가 바뀐다. 그러므로 이 논문에는 자연스럽게 목적 속성의 이전(priori) 확률분포와 이후(posteriori) 확률분포의 차이의 크기를 규칙의 중요도로 정의한다.

다음으로 중요한 부분은 이러한 정보의 양을 어떻게 정확히 측정하는 방법을 정의하고 결정하느냐이다. 본 논문에서 우리는 헬링거 함수(Hellinger divergence)를 이러한 정보를 측정하는데 사용한다. 헬링거 함수는 Beran [12]에 의해 소개되었고, 이 논문에서는 규칙의 정보 내용으로 사용하기 위해 수정하였다. 연관 규칙 $b \rightarrow a$ 에서 b 값이 주어졌다는 가정하에서 A 변수의 헬링거 함수는

$$\left(\sum_i \left(\sqrt{p(a_i)} - \sqrt{p(a_i|b)}\right)^2\right)^{1/2} \quad (1)$$

으로 정의된다. 여기서 a_i는 변수 A의 값을 의미한다. 이 값은 오직 이전 확률분포와 이후 확률분포가 동일할 때만 0이 된다. 또한 헬링거 함수는 가능한 모든 이전 확률 분포와 이후 확률분포에 대하여 연속적이다. 따라서 헬링거 함수는 이전 분포와 이후 분포의 크기 차이에 대한 측정값으로 사용할 수 있다. 그러므로 우리는 헬링거 함수를 연관규칙의 정보량으로 사용한다.

이 논문에서 우리는 헬링거 값을 약간 수정하였다. 예를 들어 $b \rightarrow a$ 의 연관 규칙에 대하여 헬링거 함수의 값을 계산하면 이 규칙의 정보 내용은 ($IC(b \rightarrow a)$로 표기) 아래와 같다.

$$IC(b \rightarrow a) = \left(\sqrt{p(a|b)} - \sqrt{p(a)}\right)^2 + \left(\sqrt{1-p(a|b)} - \sqrt{1-p(a)}\right)^2 \quad (2)$$

이때 $p(a|b)$는 $B=b$ 라는 조건하에 $A=a$ 의 조건 확률이다. 여기서 식 (2) 가 식 (1)의 정의와 다르다는 것에 주목해야 한다. 연관규칙에 있어서 법칙의 우측에는 하나의 클래스의 특정 값만이 나타나고, 해당 클래스 값은 제외한 다른 모든 값들의 확률은 $1-p(a)$에 포함된다.

또한 본 연구에서는 원래의 헬링거 수식의 제곱형태를 사용한다. 이는 1) 각 패턴의 상대적 정보 내용은 수정된 헬링거 함수에 의해 영향을 받지 않으며 2) 법칙 중요도의 다음 항목인 일반성과의 그 수치 범위상 합리적인 균형을 이루기 위해서이다.

우리가 고려해야 할 연관 규칙의 또 다른 기준은 규칙의 일반성(generality)이다. 일반성의 기본적인 아이디어는 “더 많은 데이터가 연관 규칙의 조건을 만족하면 더 유용한 법칙이 된다”이다. 본 논문에서 일반성은 다음과 같이 정의하였다.

$$G(b\rightarrow a)=\sqrt{p(b)} \tag{3}$$

본 논문에서 일반성을 표현하기 위하여 원래 확률의 제곱근 형태를 사용하는 이유는 제곱근 값이 더 정확하게 사건의 보편성을 나타낼 수 있기 때문이다. 사건의 보편성은 b 사건이 처음 나타날 때 빠르게 증가한다. 그 이후에는 사건이 충분히 발생되었을 때 그 중요도가 서서히 커진다. <그림 1>은 $p(b)$와 그것을 선형화시킨 $\sqrt{p(b)}$를 나타내는 그래프를 비교한 것이다. <그림 1>의 제곱근 함수 그래프에서 보듯이, 제곱근 값은 초기에는 급격히 증가하다가 이벤트가 많이 발생 후 다음 사건의 관찰이 덜 중요해지게 된다. 한편, $p(a)$로 표기된 보편성의 선형 함수는 현실 세계에서의 보편성의 특성과 일치하지 않는 사건의 수에 비례하여 증가한다.

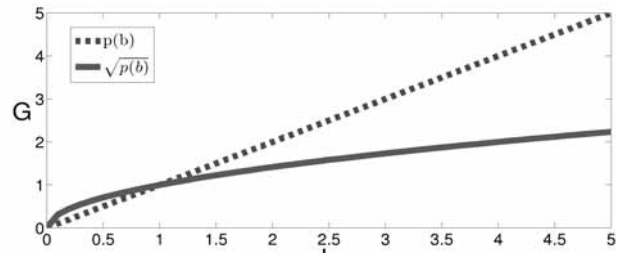

Fig. 1. Graph of $\sqrt{p(b)}$ and $p(b)$

따라서 연관 규칙의 최종적인 중요도($H(b\rightarrow a)$)는 일반성($G(b\rightarrow a)$과 규칙의 정보 내용($IC(b\rightarrow a)$)의 곱으로 나타내며 다음 식으로 표현된다.

$$\begin{aligned} H(b\rightarrow a) &= G(b\rightarrow a)\cdot IC(b\rightarrow a) \\ &= \sqrt{p(b)}\left[\left(\sqrt{p(a|b)}-\sqrt{p(a)}\right)^2+\right. \\ &\qquad \left.\left(\sqrt{(1-p(a|b)}-\sqrt{1-p(a)}\right)^2\right] \end{aligned} \tag{4}$$

이는 보편성과 정보 내용의 곱으로 구성되어 있다.

4. H measure의 속성

이 절에서는 이 논문에서 제안된 H measure 의 속성을 설명하고자 한다. 예를 들어서 $b\rightarrow a$ 규칙을 가정하고, 제안된 H measure 은 다음과 같은 속성이 있다.

속성 1 : $H(b\rightarrow a)\geq 0$.

이 속성은 규칙 중요도의 기본 속성 중 하나이다. 왜냐하면 규칙의 중요도가 음의 값을 가지는 것은 의미를 가지지 않기 때문이다. 이 속성의 증명은 식(4) 로부터 쉽게 알 수 있다.

속성 2 : $H(b\rightarrow a)=0$ 경우에만 a와 b는 독립적이다.

증명: 값 a 와 b 가 서로 독립적인 경우, $p(ab)=p(a)p(b)$ 이므로 식 (5)로부터 $H(b\rightarrow a)=0$ 이 됨을 알 수 있다. □

두 가지 변수가 서로 독립일때는 서로 원인-결과의 관계가 존재하지 않으므로 서로 독립인 변수를 사용하는 규칙의 중요도는 0 이어야한다.

속성 3 : $H(b\rightarrow a)\neq H(a\rightarrow b)$

증명: $b\rightarrow a$ 와 $a\rightarrow b$ 각 규칙에 대하여 $IC(b\rightarrow a)=IC(a\rightarrow b)$ 지만 $G(b\rightarrow a)\neq G(a\rightarrow b)$이다. 따라서 $HD(b\rightarrow a)\neq HD(a\rightarrow b)$이다. □

본 속성의 의미는 H measure 는 법칙의 원인과 결과를 구분하는 능력을 가지고 있음을 보여준다. 즉 규칙 $b\rightarrow a$ 와 규칙 $a\rightarrow b$ 는 같은 중요도를 가지지 않는다.

속성 4 : $p(a)$와 $p(b)$의 값이 고정되었을 때, $p(ab)$값이 증가하면, H measure 는 다음과 같이 동작한다.

$$H(b\rightarrow a)=\begin{cases}\searrow & \text{if } p(ab)<p(a)p(b)\\ 0 & \text{if } p(ab)=p(a)p(b)\\ \nearrow & \text{if } otherwise\end{cases}$$

↘ 및 ↗ 기호는 순차적으로 감소하고 증가하는 H measure 의 측정값을 의미한다.

증명: 식 (4) 로부터

$$\frac{\partial H(b\rightarrow a)}{\partial p(ab)}=-2\sqrt{p(a)}\left(\frac{1}{2}\right)\left(\frac{1}{\sqrt{p(ab)}}\right)-$$

$$=\sqrt{\frac{1-p(a)}{p(b)-p(ab)}}-\sqrt{\frac{p(a)}{p(ab)}} \tag{5}$$

다음의 식에서

$$D=\frac{1-p(a)}{p(b)-p(ab)}-\frac{p(a)}{p(ab)}=\frac{p(ab)-p(a)p(b)}{((p(b)-p(ab))p(ab)}$$

$$p(ab)<p(a)p(b),\ D<0,\ \frac{\partial H(b\rightarrow a)}{\partial p(ab)}<0$$

$$p(ab)=p(a)p(b),\ D=0,\ H(b\rightarrow a)=0$$

$$p(ab)>p(a)p(b),\ D>0,\ \frac{\partial H(b\rightarrow a)}{\partial p(ab)}>0 \qquad \square$$

H measure 는 a 와 b 간의 독립성으로 부터의 편차 정도에 따라서 순차적으로 증가되기 때문에 이 속성은 H measure 의 중요한 특성을 보여준다. <그림 2>는 $p(ab)$ 값에 따른 H measure 값의 변화를 보여준다.

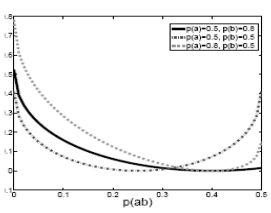

Fig. 2. H value vs. $p(ab)$

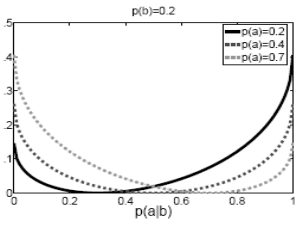

Fig. 3. H value vs. $p(a|b)$

속성 5 : $p(a)$와 $p(b)$의 값이 고정되었을때 $p(a|b)$ 값이 증가하면, H measure 는 다음과 같이 변화한다.

$$H(b\rightarrow a)=\begin{cases}\searrow & \text{if } p(a|b)<p(a)\\ 0 & \text{if } p(a|b)=p(a)\\ \nearrow & otherwise\end{cases}$$

증명: $p(b)$에 의해 속성 4 의 확률값 들을 각각 나누면 동일한 결과를 얻을 수 있기 때문에 이 속성의 증명은 간단히 알 수 있다. □

<그림 3>은 $p(b)$ 값이 0.2 인 경우에 H measure 와 $p(a|b)$의 관계를 보여주며 확률 $p(a|b)$ 는 규칙의 정확도로 해석될 수 있다.

속성 6 : $p(a|b)$와 $p(b)$의 값이 고정되었을때 $p(a)$값이 증가하면, $H(b\rightarrow a)$는 다음과 같이 동작한다.

$$H(b\rightarrow a)=\begin{cases}\searrow & \text{if } p(a)<p(a|b)\\ 0 & \text{if } p(a)=p(a|b)\\ \nearrow & otherwise\end{cases}$$

증명: 이 속성은 속성 5 의 내용을 $p(a)$ 를 기준으로 설명한 속성이다. □

속성 7 : H 는 $p(b)$ 값이 증가할수록 순차적으로 증가한다.
증명: 식(4) 를 기반으로 쉽게 성립함을 알 수 있다. □

Table 1. Accuracies of different rule measures

Dataset	*Inst.*	*Attr*	*H*	*J*	*G*	*PS*
balance	625	4	83.0	**84.0**	**84.0**	81.4
cmc	1473	9	**59.1**	31.5	*54.5*	47.9
credit	1000	15	70.0	70.0	70.0	**72.5**
crx	653	15	**85.7**	*84.5*	82.5	53.9
derm.	358	34	**63.1**	18.7	*46.3*	27.3
diabetes	768	6	**77.8**	77.7	**77.8**	66.6
flare	323	12	**59.4**	19.2	*50.1*	34.9
glass	214	7	*64.0*	**64.4**	*64.0*	59.8
haber	612	3	74.1	74.1	68.9	**75.6**
kr	3196	36	*81.7*	32.6	**86.9**	55.0
lung	27	55	**74.0**	11.1	66.6	37.0
lymph	148	18	**66.8**	**66.8**	65.5	64.1
monks	1711	6	**66.3**	63.7	63.5	65.0
nursery	12960	8	73.8	39.3	72.1	82.0

4.1 다중 규칙을 사용한 분류법

이 절에서는 연관규칙을 이용하여 클래스 값을 결정하는 방법에 대해 설명한다. 새로운 테스트 데이터가 주어지면, 규칙집합으로 부터 새로운 테스트 데이터와 일치하는 연관규칙의 집합을 수집한다.

분류학습이 효율적이기 위해서 다음과 같은 방법을 사용한다. 주어진 테스트 데이터에서 이를 만족하는 2 개의 규칙 R_g와 R_s가 있다고 가정하자. 여기서 R_g 는 R_s 에 대하여 일반적인 규칙이다. 이 경우에 우리는 좀 더 일반 규칙인 R_g 만을 고려할 필요가 있기 때문에 중요도 값과 관계없이 R_s 는 분류시에 고려하지 않는다. 따라서 테스트 데이터의 목적 클래스 값은 다음과 같이 결정된다. 클래스 라벨에 따라 그룹으로 규칙들을 나누고 각 그룹에서의 규칙들의 중요도 합을 계산한다. 최대의 중요도 합을 가진 클래스가 데이터에 대한 분류의 값이 된다.

5. 실험 평가

본 연구에서는 H measure 의 성능을 확인하기 위한 다수의 실험이 이루어졌다. H measure의 성능은 H measure 를 사용한 분류 모델의 정확도에 기반을 두어 측정된다. 실험은 두 가지 프로세스로 구성되어 있다. 첫 번 째는 일반적인 Apriori [16] 알고리즘을 사용하여 연관 규칙을 생성하고 두 번째로는 생성된 연관 규칙을 사용하여 분류기를 구축하였다. 동일한 데이터에 대하여 일반적(general)이거나 특정적인(specific) 규칙들이 공존할 때, 특정적인 규칙은 고려하지 않았다.

분류 성능은 J-measure (*J*) [8] 지니-인덱스 (*G*) [2], Piatetsky-Shapiro's (*PS*) [13] 를 포함하는 다른 중요한 rule measure 방법들과 비교하였다. 실험을 위한여 UCI [14] 에서 14개의 데이터 집합을 선택했다. 또한 데이터에서의 연속 변수는 Fayyard et. al [15] 에서 제시한 방법을 사용하여 분할하였다. 분류의 실험에는 10-fold cross validation 검사 방식을 사용했다. Apriori 알고리즘에서 사용하는 minimum support 와 minimum confidence 는 모두 10 %로 설정하였다.

<표 1> 은 이러한 rule measure 들을 이용한 분류의 정확도 결과를 보여준다. 굵게 표기된 숫자는 방법들 중에서 최고의 정확도를 의미하고, 기울임꼴 숫자는 두 번째 정확도를 의미한다. <표 1> 에 보여진 바와 같이, H measure 는 매우 좋은 성능을 보여주고 있다. H measure 는 8 개의 데이터에서 최고 성능과 2 개에서 두 번째 우수 성능을 보여주었다. 또한 1-2 번째의 성능을 보이지 않는 데이터의 경우에도 다른 measure 에 비하여 많은 정확도의 차이를

보이지 않음을 알 수 있다. 또한 다른 measure 들은 데이터의 종류에 따라서 아주 현저하게 정확도가 떨어지는 경우가 있었다. 하지만 H measure 는 이러한 현상을 보이지 않고 데이터의 종류와 특징에 관계없이 전체적으로 좋은 분류 성능을 보여주고 있었다. 이러한 결과에서 알 수 있듯이, 대부분의 경우에 H measure 를 사용한 연관법칙 기반의 분류가 다른 rule measure 들의 분류보다 더 나은 성능을 보여주었다. 또한 이는 H measure 가 분류 모델에서 다른 방법들보다 규칙의 중요도를 정확하게 나타낸다고 할 수 있다.

6. 결 론

본 연구에서, 우리는 H measure 라고 불리는 새로운 연관 규칙의 중요도를 제안한다. 그리고 이를 연관 규칙의 일반화와 결합하였다. 제안된 H measure 는 몇 가지 중요하고 흥미로운 특성을 보여준다. 학습을 통한 분류를 위한 도구로 이 measure 을 이용하였고, 현재의 다른 중요도 measure 들의 분류수행능력과 실험을 통하여 비교하였다.

실험결과 H measure 의 분류가 대부분의 경우 다른 rule measure 들에 비해 더 좋은 수행능력을 보였다. 결과적으로 본 연구에서는 H measure 는 학습을 통한 분류 분야에서 다른 rule measure 보다 연관규칙의 중요도를 더 정확하게 측정할 수 있음을 알 수 있다.

참 고 문 헌

[1] B. Liu, W. Hsu, S. Chen, and Y. Ma. Analyzing the subjective interestingness of association rules. IEEE Intelligent Systems, 15(5):47-55, 2000.

[2] R. J. Bayardo and R. Agrawal. Mining the most interesting rules. In KDD, pages 145-154, 1999.

[3] X.-H. Huynh, F. Guillet, and H. Briand. Arqat: An exploratory analysis tool for interestingness measures. In the 11th international symposium on Applied Stochastic Models and Data Analysis, pages 334-344, 2005.

[4] P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. Information Systems, 29(4):293-313, 2004.

[5] P. Lenca B. Vaillant and S. Lallich. A clustering of interestingness measures. In Proceedings of the 7th International Conference on Discovery Science, pages 290-297, 2004.

[6] P. Clark and T. Niblett. The CN2 induction algorithm. Machine Learning, 3(4):261-283, 1989.

[7] S. Jaroszewicz and D. A. Simovici. A general measure of rule interestingness. In PKDD, pages 253-265, 2001.

[8] P. Smyth and R. M. Goodman. An information theoretic approach to rule induction from databases. IEEE Transactions on Knowledge and Data Engineering, 4(4):301-316, 1992.

[9] S. Jaroszewicz and D. A. Simovici. Interestingness of frequent itemsets sing bayesian networks as background knowledge. KDD, pages 178-186, 2004.

[10] L. Wong G. Dong, X. Zhang and J. Li. Caep. Caep: Classification by aggregating emerging patterns. In Proceedings of the 2nd International Conference on Discovery Science, pages 30-42, 1999.

[11] J. Han W. Li and J. Pei. Cmar: Accurate and efficient classification based on multiple class-association rules. ICDM, pages 208-217, 2001. 14

[12] R. J. Beran. Minimum hellinger distances for parametric models. Ann. Statistics, 5:445-463, 1977.

[13] G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules, in: G. piatetsky-shapiro. In Knowledge Discovery in Databases, MIT Press., pages 229-248, 1991.

[14] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[15] U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In Int'l Joint Conference on Articial Intelligence, pages 1022-1029, 1993.

[16] R. and R. Srikant, Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994

이 창 환

e-mail : chlee@dgu.ac.kr

1982년 2월 서울대학교 계산통계학과(학사)

1988년 8월 서울대학교 계산통계학과(석사)

1994년 8월 University of Connecticut, Dept. of Computer Science(박사)

1994년 12월~1996년 2월 AT&T Bell Laboratories

2001년 1월~2001년 12월 University of Illinois, visiting professor

1996년 3월~현 재 동국대학교 정보통신학과 교수

관심분야 : 기계학습, 인공지능

배 주 현

e-mail : baegop@pusan.ac.kr

1994년 2월 부산대학교 대기과학과(학사)

1999년 2월 부산대학교 환경시스템학과(석사)

2006년 2월 부산대학교 대기과학과(박사)

2006년 3월~현 재 동국대학교 정보통신학과 강사

관련분야 : 인공지능