한국지능시스템학회 논문지 2009, Vol. 19, No. 4, pp. 525-531

인간 지식을 이용한 경험적 의사결정트리의 설계

Design of Heuristic Decision Tree (HDT) Using Human Knowledge

윤태복 · 이지형

Taebok Yoon and Jee-Hyong Lee

성균관대학교 컴퓨터공학과

요 약

데이터 마이닝(Data Mining)은 수집된 데이터로 부터 감춰진 패턴을 찾는 작업이다. 여기에서 수집된 데이터는 예측 및 추천을 위한 기반 정보로 중요한 역할을 하며, 분석 결과의 성능을 향상시키기 위해 잘못된(Missing value) 데이터를 선별하는 과정을 필요로 한다. 수집한 데이터에서 의도하지 못한 데이터를 선별하기 위한 기존의 방법은 주로 통계적이거나 단순 거리(Distance)에 기반을 둔 방법을 이용하였다. 하지만 환경 및 데이터의 특성을 고려하지 못하여, 의미 있는 데이터도 함께 분석에서 제외 될 수 있는 문제점을 가지고 있다. 본 논문은 인간의 경험적 지식을 수집된 데이터와 비교하여 가중치로 변환하고, 의사결정트리(Decision Tree)의 생성에 이용한다. 생성된 트리는 인간의 지식이 반영되어 기존의 분석 방법보다 신뢰성이 높다고 할 수 있으며, 실험을 통하여 제안하는 방법의 유효성을 확인하였다.

키워드 : 경험적 의사결정트리, 인간지식 데이터 마이닝, 이상치 데이터 감소

Abstract

Data mining is the process of extracting hidden patterns from collected data. At this time, for collected data which take important role as the basic information for prediction and recommendation, the process to discriminate incorrect data in order to enhance the performance of analysis result, is needed. The existing methods to discriminate unexpected data from collected data, mainly relies on methods which are based on statistics or simple distance between data. However, for these methods, the problematic point that even meaningful data could be excluded from analysis due that the environment and characteristic of the relevant data are not considered, exists. This study proposes a method to endow human heuristic knowledge with weight value through the comparison between collected data and human heuristic knowledge, and to use the value for creating a decision tree. The data discrimination by the method proposed is more credible as human knowledge is reflected in the created tree. The validity of the proposed method is verified through an experiment.

Key Words : Heuristic Decision Tree, Human-Knowledge Data Mining, Outlier Data Reduction

1. 서 론

데이터 마이닝(Data Mining)은 환경으로부터 수집된 데이터에서 패턴을 추출하고 의미 있는 정보를 발견하기 위하여 주로 사용된다[1]. 추출된 패턴은 미래에 실행 가능한 정보를 가지고 있어서, 예측/적응 및 지능화된 서비스에 활용 가능하다. 이때 수집된 데이터는 데이터의 특성(Continuous, Ordinal, Nominal, etc.) 및 적용하는 도메인(Ubiquitous, e-Learning, Game, etc.)에 따라, 그에 맞는 적절한 분석 과정을 거쳐, 예측 및 지능 환경 제공을 위한 지식 서비스의 기반 정보로 매우 중요하게 사용된다. 하지만, 분석을 위해 사용되는 데이터에는 분석자가 의도하지 못한 데이터가 종종 섞여 있어서 분석 및 분석 결과 해석의 어려움을 가져온다. 분석자가 의도하지 못한 데이터에는 수집 환경의 장비 오류나 결함으로 잘 못된 데이터가 수집되거나, 인간으로부터 수집할 경우, 수집 대상의 선호 및 성향의 심리적 변화, 고의적인 오류 데이터 생성 등이 이에 속한다[5]. 이러한 데이터를 선별하기 위해서는 분석 전에 전처리(Preprocessing) 과정을 통해 의미 없는 데이터를 제거(Reduction)하는 것이 일반적이다. 하지만 기존에는 통계적이거나 단순 거리(Distance)에 기반을 둔 제거 방법을 주로 이용하였다. 이런 방법들은 제거되는 데이터의 유효성 및 의미를 파악하지 못한 상태에서 실시하기 때문에 유효한 데이터도 함께 제거 될 수 있다는 문제점을 가지고 있다. 또한, 이 방법은 수집 데이터의 속성 및 환경의 특성을 고려하지 못하여 의미 있는 결과를 얻는데 어려움을 가져온다. 수집된 데이터에서 의미 있는 결과를 얻기 위해서는 도메인과 데이터의 특성을 고려한 효과적

접수일자 : 2009년 4월 6일
완료일자 : 2009년 7월 28일
이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업의 연구 결과입니다. 연구비 지원에 감사드립니다(No. 2009-0075109).
"본 논문은 본 학회 2009년도 춘계학술대회에서 선정된 우수논문입니다."

인 데이터 전처리 과정이 요구된다.

본 논문은 수집된 데이터의 의미 및 중요도를 고려하기 위해서 인간의 경험적 지식(Heuristic Knowledge)을 이용한 의사결정나무(Heuristic Decision Tree : HDT)방법을 제안한다. HDT는 인간의 경험적 지식을 이용하여 수집된 데이터를 평가하고 중요도에 따른 가중치를 계산한다. 생성된 가중치는 변형된 엔트로피(Entropy) 알고리즘에 적용하여 중요도가 높은 데이터를 선별하여 의사결정트리 생성에 반영한다. 이 방법은 의미 없는(가중치가 낮은) 데이터일수록 결과에 낮은 영향을 미치도록 하였다. 기존에는 데이터를 선별하여 의미 없다고 판단되면 제거하는 방법을 사용하였으나, 제안하는 방법은 데이터의 중요도에 따른 가중치를 부여하고, 중요한 정도를 고려하여 분석에 영향을 주는 방법을 선택하였다.

제안하는 방법의 기대효과는 다음과 같다.

- 데이터 수집 환경의 특성을 고려하여, 경험적 지식을 반영함으로써 해당 환경에 보다 적절한 결과를 얻음
- 데이터 전처리 과정에서 유효한 데이터의 손실을 최소화
- 분석 후 결과에 대한 해석이 용이하고, 사용성이 높으며, 생성된 패턴/모델/지식의 신뢰성이 높음

실험에서는 일반적인 의사결정트리 학습 방법과 경험적 의사결정트리 학습 방법을 비교 분석하였으며, 제안하는 방법의 유효성을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구에 대하여 기술하고, 3장에서는 제안하는 방법인 휴리스틱 의사결정트리 방법에 대하여 소개한다. 4장에서는 실험을 통하여 제안하는 방법의 유효성을 확인하고, 5장에서는 결론과 향후 연구로 맺는다.

2. 관련 연구

일반적으로 수집된 데이터에서 의도하지 못한 데이터 또는 알지 못하는 데이터를 불완전 데이터(Incomplete Data), 잘 못된 값(Missing Values) 또는 이상치 데이터(Outlier Data) 등으로 정의한다. 이런 데이터들을 분류하고 감소시키기 위한 연구들은 다음과 같이 진행되었다.

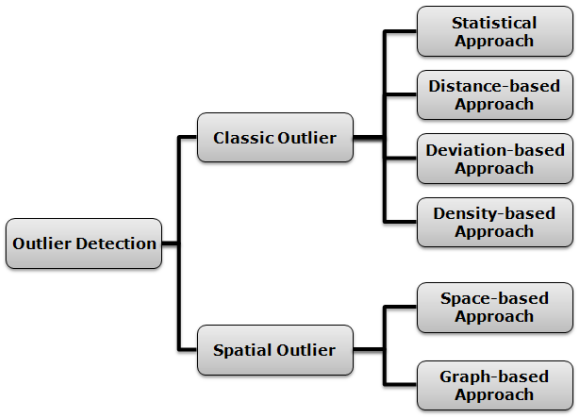

그림 1. Xi's Outlier 감소 방법 분류
Fig. 1. Xi's Outlier Reduction Methods Classification

Hwang과 Hahn[2]은 데이터 수집과정의 오류나 결함으로 인하여 모자라는 값이 발생하거나 분석자의 의도와 다른 기준으로 수집된 데이터를 불완전한 데이터라 정의하고, k-Means 클러스터링 방법을 이용하여 데이터의 결함을 예측하고 분석자의 의도에 맞게 재분류하는 연구를 하였다. Dick[3]은 이메일 스팸 제거 기술에 Incomplete Data 분류를 위해 Support Vector Machine (SVM)을 이용하여 유용한 결과를 얻었다. 또한, Xi[4]는 Outlier Reduction 방법에 대한 조사를 실시하고, 최근의 고차원 기반(High Dimension-based)에서의 이슈와 SVM 방법을 이용한 이상치 감소 방법을 소개하였다.

Kim[5]은 지능형 이러닝 시스템에서 수집된 학습자의 학습 행위 데이터에서 이상치 데이터를 감소하기 위한 연구를 수행하였다. 이 논문에서는 군집화 방법을 이용하여 이상치를 감소시키고 학습자 모델의 성능이 향상된 것을 실험을 통하여 확인하였다. Lee와 Choi[6]는 방대하고 불분명한 자료 및 정보를 해석하는데 있어서 여러 속성을 이용한 분류화 및 근사화를 효과적으로 제공하는 러프집합이론(Rough Sets theory) 소개하고, 위스콘신 대학병원의 유방암 관련 데이터를 분석에 이용하여, 뉴로-퍼지, C4.5 와 제안하는 방법인 계층적 러프분류기법을 이용하여 비교 실험하였다. M¨uller[7]는 고차원 데이터에서 이상치 데이터에 대한 순위를 결정하는 OutRank 방법에 대하여 소개하였고, Jiang와 An[8]은 이상치 데이터를 선별하기 위해 군집화 방법을 이용하였다. Zhang과 Lu[9]는 잘 못된 수집데이터를 선별하기 위해 베이시안 네트워크를 이용하였으며, Zheng[10]은 웹 사용 마이닝을 위한 모델 생성에 불완전 데이터 문제를 처리하기 위한 방법을 소개하였다. 이처럼 알지 못하는 데이터를 선별하고 분류하기 위한 다양한 방법이 연구되고 있으나, 기존의 방법들은 데이터의 의미를 고려하지 않고, 선별/제거하는데 목표를 두고 있다. 수집 데이터의 의미를 고려한 데이터 선별 및 분류 방법이 요구된다.

3. 경험적 의사결정트리 (Heuristic Decision Tree : HDT)

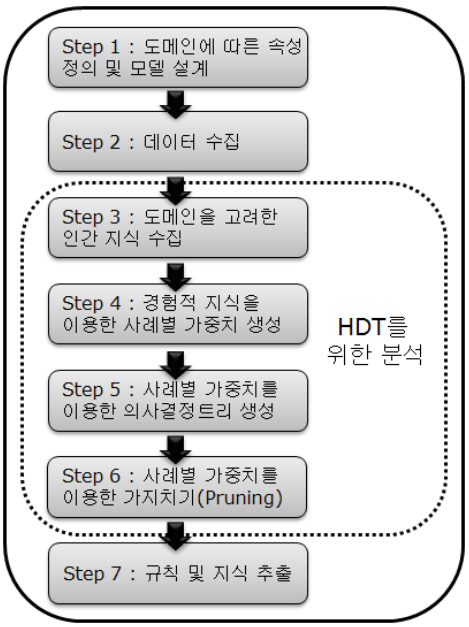

그림 2. HDT를 위한 단계별 전체 작업 흐름도
Fig. 2. Work flowchart for HDT

경험적 의사결정트리(HDT)는 인간의 지식을 이용하여 수집데이터에 가중치를 부여하고, 그 가중치는 의사결정트리 생성 과정에 영향을 미친다. 생성된 트리는 경험적 지식이 반영되어 보다 의미 있는 정보를 가지고 있으며, 다음은 그림은 제안하는 방법인 HDT를 위한 단계별 작업 흐름도이다.

일반적인 데이터 마이닝 작업은 설계, 수집, 가공, 분석, 해석의 과정을 거친다. 제안하는 방법은 일반적인 방법과 유사하나 가공과 분석과정에 인간의 지식을 반영하기 위해 경험적 지식 수집(Step 3)과 사례에 대한 가중치 생성(Step 4) 그리고 사례별 가중치(Step 5,6)를 분석에 적용하기 위한 과정이 추가 되었다.

Step 1에서 도메인에 따른 속성의 정의 및 분석 후에 얻고자 하는 모델을 설계한다. Step 2에서는 Step 1에서 정의된 속성에 따라 데이터를 수집한다.

표1은 Step 2의 한 예이다. n개의 속성들(Attributes)을 가지고, 클래스(Class)는 Y와 N을 가지며, 사례(Instance) m개를 나타낸 경우이다. 예를 들어 사례 2(I_2)의 경우 A_1이 12, A_2는 32, A_3는 27, 그리고 A_n은 19의 값을 가지며, 그 때 클래스는 Y이다.

표 1. 분석을 위한 수집데이터 예.
Table 1. An example of collected data for analysis.

	A_1	A_2	A_3	...	A_n	C
I_1	20	12	23	...	23	Y
I_2	12	32	27	...	19	Y
I_3	25	12	19	...	28	N
...	...	...	...	...	...	...
I_m	23	31	23	...	23	Y

3.1 도메인을 고려한 경험적 지식 수집

분석가는 도메인에 대하여, 사전에 정의된 속성에 따라 인간의 경험적 또는 전문가적 지식을 수집한다. 도메인의 특성 및 수집 데이터의 특성을 고려하여 데이터를 자동으로 선별하는 작업에서 인간의 경험적 지식을 이용하지 않는 것은 매우 어려운 일이다. 표 2는 n개의 속성 A에 대한 j개의 경험적 지식 H를 수집한 예를 보여주고 있다. 또한, 정의된 경험적 지식의 신뢰도(Confidence degree: CF)를 0~1값으로 인간 스스로 부여함으로써, 확실하지 않은 지식도 수렴할 수 있도록 하였다.

표 2. 분석을 위한 경험적 지식의 예.
Table 2. An example of Human-knowledge for analysis.

	A_1	A_2	A_3	...	A_n	C	CF
H_1	23	12	23	...	22	Y	0.9
H_2	21	17	N/A	...	20	N	0.8
H_3	15	N/A	24	...	25	N	0.4
...	...	...	...	...	...	...	...
H_j	26	21	N/A	...	22	Y	0.4

예를 들어, H_1은 속성 A_1에 대하여 23, A_2는 12, A_3는 23 그리고 A_n은 22 일 때, 클래스는 Y 일 것이라는 경험적 지식이다. 또한, 이때 제시한 H_1의 신뢰도는 0.9를 나타내고 있다. 수집된 경험적 지식은 표 1의 수집된 데이터와 연산을 통하여 가중치 테이블을 생성하게 된다.

3.2 경험적 지식을 이용한 사례별 가중치 생성

수집된 데이터(표 1)와 인간의 경험적 지식(표 2)을 이용하여 그림 3의 우측 테이블과 같은 가중치 테이블을 생성한다.

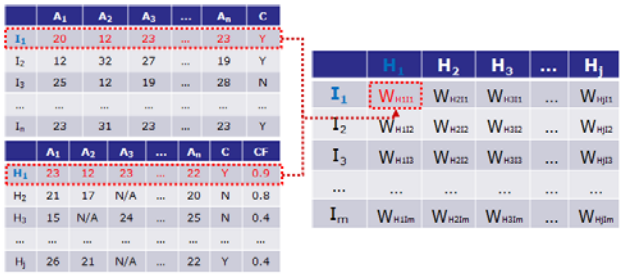

그림 3. 수집데이터와 경험적 지식을 이용한 가중치 테이블 생성 예.
Fig. 3. An example of weighted-table using collected data and heuristic-knowledge.

수집된 데이터의 사례 I와 경험적 지식 H를 이용한 가중치 계산은 수식(1)을 이용한다. 수식(1)에서 n은 전체 속성의 개수를 나타내며, I와 H의 i번째 속성 값은 유사할수록 2에 가까운 값을 얻을 수 있도록 하였다. 수집 데이터와 인간의 경험적 지식이 일치 할 경우 1의 값을 얻으며, 값이 멀어질수록 0에 근접한 값을 얻는다. 또한, 경험적 지식의 신뢰도(CF)를 수식에 포함하여 신뢰 정도에 따라 다른 결과가 나올 수 있도록 하였다. Ai_{max}는 Ai의 최대값, Ai_{min}은 Ai의 최소값을 의미한다. 또한 ImAi는 속성 i의 m번째 사례, HjAi는 속성 i의 j번째 인간지식을 의미한다.

$$W_{I_m,H_j} = \sum_{i=1}^{n}\left(\left(1-\frac{\left|I_m A_i - H_j A_i\right|}{A_{i_{max}} - A_{i_{min}}}\right)\cdot\frac{1}{n}\right)\cdot CF \qquad (1)$$

3.3 사례별 가중치를 이용한 의사결정트리 생성

사례별 가중치를 계산하여 표 3과 같은 결과를 얻었다고 가정할 때, 각 사례의 가중치 중에서 가장 큰 값을 H_{max}라고 표현하고, H_{max}를 수집 데이터에 포함한다. 표 4는 가중치 테이블에서 H_{max}를 이용하여 수집 데이터에 HDT_{weight}를 추가하여 다시 표현하였다.

표 3. 생성된 가중치 테이블 예.
Table 3. An example of weighted-table.

	H_1	H_2	H_3	...	H_n	H_{max}
I_1	0.8	0.4	0.6	...	0.5	0.8
I_2	0.2	0.4	0.1	...	0.4	0.4
I_3	0.5	0.7	0.4	...	0.7	0.7
...	...	...	...	...	...	...
I_n	0.8	0.7	0.9	...	0.5	0.9

표 4. 휴리스틱 가중치가 포함된 수집 데이터.
Table 4. Collected data included heuristic weight

	A_1	A_2	A_3	...	A_n	C	HDT_{weight}
I_1	20	12	23	...	23	Y	0.8
I_2	12	32	27	...	19	Y	0.4
I_3	25	12	19	...	28	N	0.7
...	...	...	...	...	...	...	...
I_m	23	31	23	...	23	Y	0.9

전문가 가중치를 이용한 의사결정트리를 생성하기 위해서는 기존과 다른 방법을 이용하여야 한다. 데이터 분석에 HDT_{weight}를 적용할 수 있어야 한다. 다음은 변형된 엔트로피 공식을 위한 설명이다.

우선 의사결정트리 알고리즘에서 정의하는 엔트로피공식을 소개한다. 일반적인 의사결정트리 학습 방법을 위한 엔트로피는 수식(2)와 같다.

$$Entropy(S) = -\sum_{c=1}^{T_c} P_c \log_2 P_c \qquad (2)$$

S는 사례들의 집합이고, T_c는 사례들이 속하는 클래스의 총 개수이며, P_c는 S중에서 클래스 c에 속하는 사례들의 비율이다. 즉 S에 n개의 사례가 있고 이중 n_c개가 있다면 $P_c=n_c/n$가 된다. 그리고 $0\log_2 0=0$로 정의한다. S중에서 c에 속하는 것이 하나이면 P_c =1/n이고 두 개이면 2/n이 되므로 S의 각 사례가 S의 엔트로피에 미치는 영향은 1/n에 비례한다고 할 수 있다. 즉 i번째 사례가 갖는 가중치를 W(i)로 표시하고, 모든 i에 대하여 W(i)=1이라고 한다면, P_c는 다음과 같이 다시 나타낼 수 있다. c(i)는 i번째 사례가 속하는 클래스이다.

$$P_c = \frac{\sum_{c(i)=c} W(i)}{\sum_{i=1}^{n} W(i)} \qquad (3)$$

기존의 의사결정트리 알고리즘은 모든 사례의 가중치 값을 모두 1로 계산한다. 하지만, 제안하는 방법의 경우 각 사례에 대하여 경험적 지식을 이용한 가중치(HDT_{weight})가 부여되어 있다. HDT_{weight}를 의사결정트리 생성에 반영하기 위해 아래와 같이 새로운 엔트로피를 산출하는 공식은 수식(4)와 같다.

$$Entropy(s) = -\sum_{c=1}^{T_c} \frac{\sum_{c(i)=c} HDT_{weight}(i)}{\sum_{i=1}^{n} HDT_{weight}(i)} \log_2 \frac{\sum_{c(i)=c} HDT_{weight}(i)}{\sum_{i=1}^{n} HDT_{weight}(i)} \qquad (4)$$

3.4 사례별 가중치를 이용한 가지치기(Pruning)

가지치기는 분류오류를 크게 할 위험이 높거나 부적절한 추론규칙을 가지고 있는 트리의 가지(Branch)를 제거하는 것을 말한다. 생성된 트리 구조가 지나치게 복잡한 구조를 가질 경우 새로운 사례에 대하여 예측오차가 크게 나올 수 있다. 이런 경우를 과잉 학습(Over fitting)되었다고 이야기하며, 가지치기를 통하여 예측 오차를 줄이는 효과를 볼 수 있다.

가지치기는 의사결정나무를 형성하며 분리 할 때 정지규칙을 적용하여 가지치기를 수행하는 사전 가지치기(Pre-Pruning)와 의사결정나무를 생성한 후 상위노드와 하위노드의 에러율을 비교하여 가지치기를 결정하는 사후 가지치기(Post-Pruning)로 구분되어진다. 본 논문에서는 사후 가지치기 방법 중에 Rule Post-Pruning(RPP) [11] 방법을 이용하였다. RPP 방법은 생성된 트리의 모양을 직접 변형하는 것이 아닌, 의사결정트리로부터 얻은 If-then형태의 규칙을 가지치기 데이터(Growing Data)를 이용하여 오류율이 클 경우 제거하는 방식을 이용한다(그림 4).

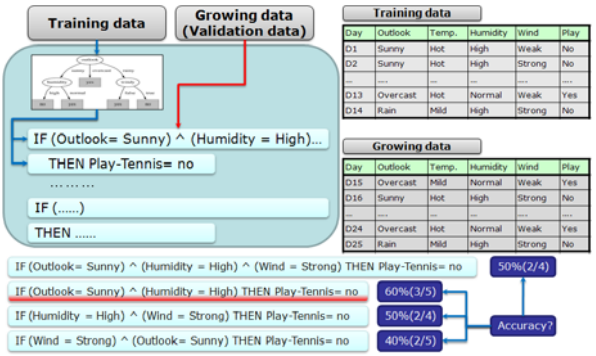

그림 4. 의사결정나무에서 학습과 가지치기
Fig. 4. Learning & Pruning in Decision Tree Method

일반적으로 수집된 데이터는 학습 데이터(Training Data)와 검증 데이터(Test Data)로 분류하여 실험을 한다. 학습 데이터는 분석과정을 거쳐 모델 생성에 사용되며, 검증 데이터는 생성된 모델의 예측 오차를 측정하기 위해 사용된다. 앞서 설명한 RPP와 같은 사후 가지치기를 위해서는 학습 데이터에서 일부를 가지치기 데이터(Growing Data 또는 Validation Data)로 분류한다. 가지치기 데이터는 학습 데이터를 이용하여 생성된 모델에 대입하여 오류 정도를 측정하고 가지치기에 판단에 이용된다. 학습 데이터로부터 생성된 트리는 If-then 형태의 규칙으로 표현한다. 전통적인 가지치기 방법은 생성된 규칙이 가지치기 데이터에서 일치하는 정도를 단순 일치 정도에 따라 이용한다. 즉, 일치하는 사례 4가지가 존재할 때, 그중에서 2가지가 동일한 클래스에 속한다고 하며, 50%의 에러율을 가진다고 이야기 한다(그림 5). 하지만, 본 논문에서는 일치하는 사례에 대하여 단순 횟수가 아닌 표 4의 HDT_{Weight} 값을 이용하여 오류 정도 계산에 반영한다. 어떤 사례에 대하여, 속성 값이 일치하는 개수가 5개이고 이중에서 3개가 일치한다면 정확도는 (3/5*100) 60%이다. 하지만, HDT_{Weight} 값을 이용한다면 가중치의 값의 따라 다양한 정확도가 계산된다. 만약 일치하는 3개의 속성이 HDT_{Weight} 가중치가 모두 0.1 일 경우 ((0.1*3)/5*100) 6%이고, 가중치가 모두 1.5 일 경우 ((1.5*3)/5*100) 90%의 정확도를 가지게 된다.

4. 실 험

제안하는 방법의 실험을 위해 UC Irvine Machine Learning Repository[12] 데이터를 이용하여 검증하였다. 실험에 사용한 데이터는 Iris Data Set, Breast Cancer Wisconsin (Diagnostic) Data Set, Abalone Data Set이며 실험 방법은 그림 6과 같다.

그림 5. Rule Post-Pruning의 예
Fig. 5. For Example of Rule Post-Pruning

각 실험 데이터는 학습 데이터 50%, 검증 데이터 30%, 가지치기 데이터 20%로 분류하고, 각 실험환경에 따른 경험적 지식을 이용하여 경험적 가중치 테이블을 생성하였다. 제안하는 방법의 유효성을 기존 방법과 비교 검증하기 위해 3번의 다른 조건에서 실시하였다. 첫 번째는 의사결정나무 학습 방법을 이용하고, 검증 데이터를 이용하여 에러율을 측정하였다. 두 번째는 의사결정나무 학습 후, RPP를 이용하여 가지치기를 하고 검증 데이터를 이용하여 에러율을 측정하다. 마지막으로 제안하는 방법은 HDT방법을 이용하여 학습시키고, HDT_{Weight} 를 이용하여 가지치기한 후의 검증 데이터를 이용한 에러율을 측정하였다. 3가지 다른 조건의 실험은 5회 Cross-Validation 방법으로 실시하였다.

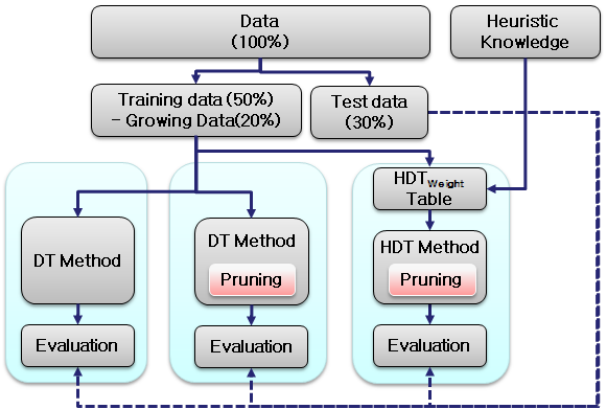

그림 6. 실험방법을 위한 전체 흐름도
Fig. 6. Work-flow for Experiment

4.1 Iris Data를 이용한 실험

Iris data는 꽃잎과 꽃받침의 길이와 폭에 따른 꽃 종류를 분류하는 것을 목표로 한다. 4개의 속성(Attributes)을과, Iris Versicolour와 Iris Virginica 두 가지 클래스(Class)를 가지고 있다. 전체 사례(Instance)의 개수는 100개이며, 이중 50개는 학습 데이터, 30개는 검증 데이터, 20개는 가지치기 데이터로 사용하였다. 그림 7는 단순한 의사결정나무 학습(Normal DT) 후의 에러율, 의사결정 나무 학습과 가지치기(Pruning DT) 후의 에러율, 제안하는 방법인 경험적 지식을 이용한 의사결정나무(Heuristic DT) 학습 후의 에러율을 나타내고 있다.

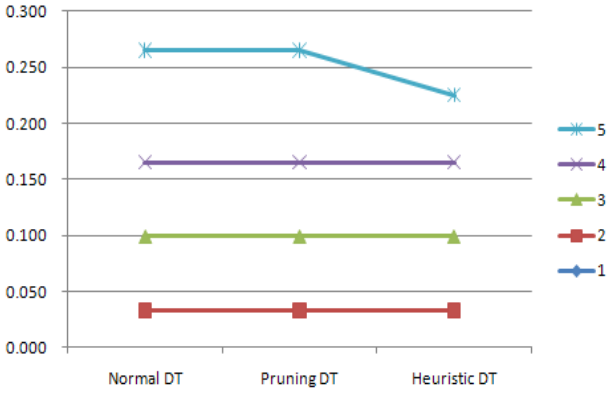

그림 7. Iris Data의 실험 결과
Fig. 7. Experiment Results of Iris Data

4.2 Breast Cancer Wisconsin (Diagnostic) Data를 이용한 실험

Breast Cancer Diagnostic Data는 세침 흡인세포검사를 통하여 유방암에 대한 악성 및 양성을 나타내는 데이터이다. 속성에는 세포 이미지의 둘레, 면적, 부드러운 정도 등에 따라 10가지를 가지며, 클래스는 악성과 양성 두 가지로 나뉜다. 본 실험에서는 악성 데이터 200개, 양성 데이터 200개, 전체 400개의 데이터를 사용하였다. 그림 8은 3가지 다른 조건에서의 실험 결과이다.

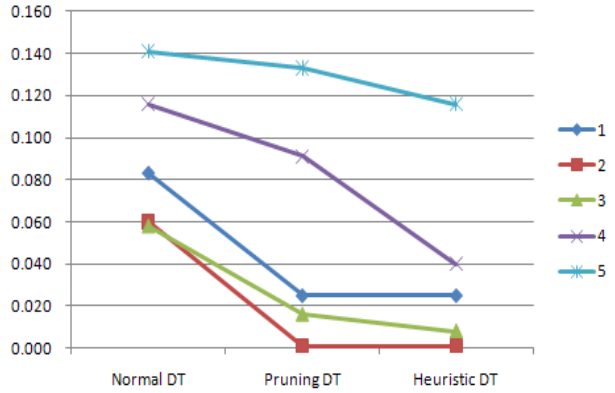

그림 8. Breast Cancer Data의 실험 결과
Fig. 8. Experiment Results of Breast Cancer Data

4.3 Abalone Data Set를 이용한 실험

Abalone Data는 전복의 상태에 따른 나이를 나타내는 데이터이다. 속성에는 전복의 길이, 무게, 지름, 내장무게, 표면 무게 등 7가지의 속성과 클래스는 1~29년 사이의 값을 갖는다. 본 실험에서는 10년생 이상과 이하, 두 가지 클래스로 분류하고 사용하였다. 분석에 사용된 데이터는 10년생 이하 1000개, 10년생 이상 1000개, 총 2000개를 이용하였다. 그림 9는 실험결과를 나타내고 있다.

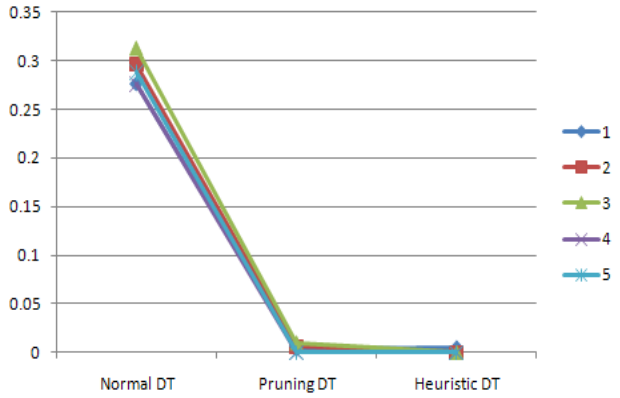

그림 9. Abalone Data의 실험 결과
Fig. 9. Experiment Results of Abalone Data

5. 결론 및 향후 연구

본 논문은 인간의 경험적 지식을 데이터 마이닝 분석에 적용하여 보다 의미 있는 정보를 얻는 방법을 제안하였다. 환경에 따라 정의된 속성에 대한 경험적 지식을 수집하고, 수집된 데이터의 각 사례별 가중치를 계산하였다. 가중치는 변형된 엔트로피와 가지치기에 적용하여 의사결정트리 생성에 인간의 경험적 지식이 반영되어 생성될 수 있도록 하였다. 제안 사항은 이상치(Outlier data) 및 불완전 데이터(Incomplete data)에 적절하게 대응할 수 있는 방법으로 데이터의 손실 없이, 분석 결과를 얻을 수 있다.

표 5. 실험 결과 비교
Table 5. Comparison of Experiment Result

	Error Rate		
	Normal DT	Pruning DT	Heuristic DT
Iris	0.053	0.053	0.045
Breast Cancer	0.092	0.053	0.038
Abalone	0.290	0.004	0.001
avg.	0.145	0.037	0.028

제안하는 방법의 검증을 위한 실험에서는 3가지 다른 도메인 데이터를 이용하여 일반적인 의사결정나무 학습 방법, 의사결정 나무 방법과 가지치기 방법 그리고 제안하는 방법인 경험적 의사결정나무 방법에 대한 에러율을 측정하였다. Iris 데이터 실험의 경우 전체 데이터의 개수가 100개 였으며, 이 데이터는 다시 학습 데이터, 검증 데이터, 가지치기 데이터로 나누어 사용하였다. 에러율이 크게 개선되지 않은 이유는 데이터의 개수가 너무 작았기 때문이다. 그에 반해 유방암 관련 데이터와 전복 데이터의 분석결과는 제안하는 방법인 경험적 의사결정나무가 기존의 방법에 비하여 에러율이 감소한 것을 확인 할 수 있다. 데이터 분석 앞서 인간의 경험적 지식을 수집하는 것은 쉽지 않은 작업니다. 향후 연구로는 인간의 경험적 지식을 보다 효과적으로 처리하고 분석할 수 있는 방법이 필요하겠다.

참 고 문 헌

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth,"Knowledge Discorvery and Data Mining : Towards a Unifying Framework", *Proc. KDD-96*, 1996.
[2] Sun-Young Hwang, H. E. Hahn,"Pre-Adjustment of Incomplete Group Variable via K-Means Clustering", *Journal of Korea Data & Information Science Society*, Vol. 15, No. 3, 2004.
[3] Uwe Dick, Peter Haider, Tobias Scheffer,"Learning from Incomplete Data with Infinite Imputations", *Proceedings of the 25th International Conference on Machine Learning*, 2008.
[4] Jingke Xi, "Outlier Detection Algorithms in Data Mining," *IEEE Second International Symposium on Intelligent Information Technology Application*, 2008.
[5] Yongse Kim, Taebok Yoon, Heonjin Cha, Youngmo Jung,Eric Wang and Jee-Hyong Lee, "An Outliers Analysis of Learner's Data based on User Interface Behaviors", *Proc. 7th IEEE Int'l. Conf. Advanced Learning Technologies (ICALT)*, 2007.
[6] Chul-Heui Lee, Sang-Chul Choi,"Discovering Classification Knowledge for Data Mining using Rough Sets and Hierarchical Classification Structure," *Journal of Telecommunication and Information*, Vol. 5, pp.79~85, 2001.
[7] Emmanuel M¨uller, Ira Assent, Uwe Steinhausen, Thomas Seidl,"OutRank: ranking outliers in high dimensional data",*International Conference on Data Engineering (ICDE) Workshop 2008*.
[8] Sheng-yi Jiang, Qing-bo An,"Clustering-based Outlier Detection Method",*Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, 2008.
[9] Hongwei Zhang, Yuchang Lu,"Learning Bayesian network classifiers from data with missing values", Proceedings. *IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering(TENCON '02)*, 2002
[10] Zhiqiang Zheng,"On an incomplete data problem in modeling: Evidence from Web usage mining and a general purpose solution", Dissertation, *University of Pennsylvania*, 2003.
[11] Trong Dung Nguyen, Tu Bao Ho, Hiroshi Shimodaira, "A Scalable Algorithm for Rule Post-pruning of Large Decision Trees", *Proceedings of the 5th Pacific-Asia Conference on Knowledge*, 2001.
[12] "http://archive.ics.uci.edu/ml/index.html", *UC Irvine Machine Learning Repository Website*.

저 자 소 개

윤태복(Taebok Yoon)
2001년 : 공주대학교 전자계산학과(학사)
2005년 : 성균관대학교 컴퓨터공학(석사)
2007년 : 성균관대학교 컴퓨터공학 (박사수료)

관심분야 : 사용자 모델링(User Modeling), 게임 인공지능 (Game A.I.)
Phone : 031-290-7987
Fax : 031-299-4637
E-mail : tbyoon@skku.edu

이지형(Jee-Hyong Lee)
1993년 : 한국과학기술원 전산학과(학사)
1995년 : 한국과학기술원 전산학과(석사)
1999년 : 한국과학기술원 전산학과(박사)
2002년～현재 : 성균관대학교 정보통신 공학부 부교수

관심분야 : 지능시스템, 기계학습, 온톨로지
Phone : 031-290-7154
Fax : 031-299-4637
E-mail : jhlee@ece.skku.ac.kr