Proceedings of KFIS Autumn Conference 2007 Vol. 17, No. 2.

전문가 의견을 반영하는 향상된 의사결정나무의 엔트로피 기법

Decision Tree Algorithm with Improved Entropy Using an Expert Opinion

박선빈, 김동문, 윤태복, 이지형
성균관대학교 정보통신공학부

SunBin Bak, DongMoon Kim, TaeBok Yoon, Jee-Hyong Lee
School of Information and Communication Eng., SungKyunKwan University, Korea

E-mail : {sam445, skyscrape, tbyoon}@skku.edu, jhlee@ece.skku.ac.kr

요 약

최근 데이터의 양이 많아지고 다양해짐에 따라서 데이터를 활용하기 위한 데이터 마이닝에 관한 관심이 증대되고 있다. 데이터 분석을 위한 수집 데이터에는 수집 과정에서 분석가가 원치 않은 데이터 잡음이 발생하는 경우가 있고 그 데이터가 다른 데이터들과 같은 가중치로 데이터 마이닝에 반영되는 경우 예상과 다른 결과를 얻을 수 있다. 따라서 데이터 분석 시 데이터와 전문가 의견이 고려된 데이터 엔트로피(Entropy)를 사용하여 잡음 데이터를 다룰 필요가 있다.

본 논문에서는 전문가의견을 이용한 전문가 의견 목록을 만들고 이를 데이터와 비교하여 유사한 정도에 따라 각 데이터에 가중치를 부여한다. 그리고 이 데이터를 활용한 의사결정나무(Decision Tree)를 사용하여 기존 데이터를 이용한 의사결정나무 보다 데이터 잡음의 영향을 줄이는 방법을 제안한다. 제안한 방법은 학습자의 학습 활동에서 수집된 학습 행위 데이터를 사용하여 실험하였다.

Key Words : Expert Opinion, Decision Tree, Preprocessing, Entropy

1. 서론

여러 기업과 대형 마켓들은 많은 거래 실적을 통해 방대한 고객 데이터들을 가지고 있다. 그리고 여러 고객들을 상대하므로 그 대상들의 특성에 따라 다양한 마케팅 전략을 세운다. 그러므로 소비자의 구매형태를 예측하거나 변수간의 인과관계를 분석하여 판매를 촉진하는 마케팅 기법이 요구된다[3]. 같은 데이터들을 이용하더라도 그 대상들에 따라 데이터 활용 방법은 다를 수 있다. 예를 들어, A란 지역의 매장 고객들과 B란 지역의 매장 고객들의 특성이나 취향은 다를 수 있다. 그러므로 같은 데이터를 사용하더라도 A지역 매장에 사용 될 때와 B지역 매장에 사용 되어질 때 데이터의 가치는 데이터의 속성 값에 따라 각기 다를 수 있다. 그러나 기존의 방법은 하나의 데이터가 모두 같은 가중치를 갖고 사용된다. 따라서 데이터 속성 값들에 따른 데이터의 중요도를 평가하는 개별적인 데이터 전처리 단계가 요구될 수 있다.

또한 데이터마이닝의 필수 요소는 신뢰도가 높고 충분한 자료이다. 신뢰도 높은 충분한 자료가 정확한 예견을 가능하게 하기 때문이다[2]. 그러나 수집 과정에서 분석가가 원치 않은 데이터 잡음이 생길 수 있고 이런 데이터는 분석가의 예견 능력을 떨어뜨릴 수 있으므로 최적의 결과를 산출할 수 있는 자료가 요구된다.

본 연구에서는 데이터에 대한 전문가의 의견과 그와 유사한 데이터들에 높은 가중치를 적용하여 다른 데이터들과 차별을 두는데 목적을 둔다. 또한 의견들과 많은 차이를 보이는 데이터는 비현실적인 데이터일 가능성이 있으므로 적은 가중치를 받아 잡음 데이터로 처리될 수도 있다. 예를 들어, A 지역에 소비자에 대한 데이터 마이닝을 할 시에는 A 지역의 소비자 성향에 맞는 데이터 속성들이 제시되고 제시된 조건들과 차이가 큰 데이터들은 적은 가중치를 받아 잡음 데이터로 전 처리 되는 방법이다.

최근에 데이터의 시간 속성에 따라 최근의 데이터 자료에 좀 더 많은 의미(가중치)를 부여하고 오래된 자료일수록 점점 작은 비중을 두게 하는 데이터 마이닝 연구가 진행되어 있다[1][3]. 하지만 데이터의 중요도는 시간 속성뿐 만 아니라 다른 속성들도 고려될 수 있다. 본 논문은 데이터의 모든 속성을 고려한다는 점이 기존의 연구와 구별된다.

2. 엔트로피 추출

2.1 방법

의사결정나무(Decision Tree)를 사용하는 방법에

는 여러 알고리즘이 있다. 그 중 ID3 알고리즘에서 사용되는 Entropy공식은 다음과 같다.

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

- **S** : 전체 사례들의 집합
- **Tc** : 사례들이 속하는 클래스의 총 개수
- **Pc** : **S**중에서 클래스 **C**에 속하는 사례들의 비율

만약 S에 n개의 instances가 있으면 C class를 Nc라 표현할 수 있다. 그리고 Pc=Nc/n 이다. 그리고 C class에 속하는 instances가 1개 이면 Pc=1/n 이다. 따라서 각 instances가 갖는 비중은 1/n로 같다[1]. 이와 같이 기존의 데이터 마이닝에서는 데이터베이스에서 데이터를 추출하여 유용한 패턴 등을 발견하고 각각의 데이터들을 모두 동일한 비율로 활용되었다.

본 논문에서는 각 Instance의 가중치를 구하기 위해 전문가 의견을 이용한다. 전문가 의견은 Instance의 속성들에 값을 대상에 대하여 현실적이고 가치 있는 대안을 몇 가지 선정하여 List로 정해둔다. 다양한 Class 에 맞는 있을 수 있는 현실적인 조건들이 정해지고 이들 조건들을 실제 데이터와 비교하게 된다.

데이터 속성 값은 명목형(Nominal), 순서형(Ordinal), 연속형(Continuous) 3가지로 구분 할 수 있다. 본 논문에서는 데이터속성의 첫 번째와 세 번째, 즉 명목형과 연속형 데이터 타입을 이용한다.
명목형 데이터는 동일 시 에만 의미가 있다고 가정을 한다.

예) 금==금, 일!=월.

연속형 데이터는 숫자들 차이로 비교를 한다.

예) 10-4=6.

Instance A는 이름, 나이, 성별, 월 수입의 4가지 속성과 Class(T,F)가 있다고 가정하자. 가치 있는 데이터 속성은 이름을 제외한 것이라 할 수 있다. 그리고 다음과 같은 전문가 의견 리스트가 있다.

- **ExpertOpinionList**

List 1) 나이1, 성별1, 월 수입1, Class1
List 2) 나이2, 성별2, 월 수입2, Class2
List 3) 나이3, 성별3, 월 수입3, Class3

본 논문에서는 각 속성의 가중치 범위를 $0 \le EachAttributeDomain \le 1/k$으로 각 속성이 같은 가중치 범위를 같게 하겠다.

n을 ExpertOpniomList의 수라고 하고, t는 ExpertOpniomList의 index라고 하자. k는 데이터 속성의 수이다. i를 Instance의 속성 Index라 하고, $W_A(i)$를 Instance와 List의 각 속성 비교 값이라 하면 $W_L(t)$는 Instance와 List를 비교하여 얻은 값이다($\sum W_A(i)$, i=1,2,...k). 따라서 p를 Instance의 index라 하면 각 Instance의 WeightedInstanceValue(p)는 Max($W_L(n)$)로 얻을 수 있다. 즉, Instance A를 List1~3과 비교하여 비교도 $W_L(t)$이 가장 높은 값을 뽑아서 가중치 적용을 한다.

명목형 데이터인 경우 List(t)와 비교하여 같은 경우 T, 다른 경우 F 를 받는다.$W_A(t)$=T =1/k or F=0 값을 갖는다. 연속형 데이터인 경우 H(i)는 속성의 index이고 H(i)의 Max, Min 값의 차이로 1/k를 나누어 H(i)의 각 값별(속성의 1단위 값) 차이를 수치화한다.

- 알고리즘 단계

[단계1] 초기화
- ExpertOpinionList를 만든다.
- 각 Attribute별 구간 설정 및 수치화 한다.

[단계2] 가중치 계산
- 데이터와 ExpertOpinionList를 비교하여 각 List에 대하여 비교도($W_L(t)$)를 구한다.
- $W_L(t)$값 중 가장 큰 값이 해당 Instance의 가중치이다.

[단계3] 가중치 적용
- 데이터에 [단계2]에서 얻은 가중치를 적용 한다.

2.2 예제

간단한 예제를 통해 본 연구에서 제안한 알고리즘을 살펴보자. 먼저 데이터베이스에는 다음과 같은 자료가 있다고 가정한다.

표 1. Training Data on the PlayTennis

날씨	온도	습도	바람	Class
sunny	25	67	weak	no
sunny	24	70	strong	no
overcast	25	66	weak	yes
rainy	21	80	weak	yes
rainy	18	72	weak	yes
rainy	19	76	strong	no
overcast	19	65	strong	yes
sunny	20	64	weak	no
sunny	21	60	weak	yes
rainy	20	76	weak	yes
sunny	18	61	strong	yes
overcast	19	69	strong	yes
overcast	23	65	weak	yes
rainy	20	80	strong	no

이 데이터는 해당 시간(데이터 수집시간)에 주어진 조건에서 테니스를 (친다/치지 않는다)를 나타내는 데이터이다. ExpertOpinionList는 다음과 같다.

- **ExpertOpinionList**

1) Sunny, 21, 65, Weak, Yes
2) Sunny, 27, 75, Strong, No
3) Rainy, 22, 77, Strong, No

이제 한 Instance를 예로 본 논문이 제안하는 방법을 적용해보자.

날씨	온도	습도	바람	Class
rainy	20	80	strong	no

Each Attribute = 1/k = 1/5 = 0.2

1) 날씨는 그날의 날씨 정보이다. Sunny, Overcast, Wind 값을 갖는다. 명목형 데이터이고 같으면 T, 다르면 F 를 갖는다.

Outlook=T(0.2)/F(0)

2) 온도는 그날의 온도이다. 연속형 데이터이고 List의 값과 근접 할수록 비교도가 높아 더 높은 가중치를 얻는다.

18≤Temperature≤25
Temperature Value 1 = 0.2/7 = 0.029

3) 습도는 그날의 습도이다. 연속형 데이터이고 List의 값과 근접 할수록 비교도가 높아 더 높은 가중치를 얻는다.

61≤Humidity≤80
Humidity Value 1 = 0.2/19 = 0.011

4) 바람은 그날의 바람의 세기이다. Strong, weak 두 단계이다. 명목형 데이터로 같으면 T, 다르면 F 를 갖는다.

Windy=T(0.2)/F(0)

5) Class는 테니스를 치면 Yes, 치지 않으면 No를 갖는 Class Value 이다. 명목형 데이터로 같으면 T, 다르면 F를 갖는다.

Class=T(0.2)/F(0)

Data 와 ExpectOpinionList 1을 비교하면 (F,-1,-15,F,F)를 얻을 수 있다. 차이 값이므로 음수를 절대 값으로 변환하면 (F,1,15,F,F) 를 얻을 수 있다.

각 속성별 Value 값 :
(0.2, 0.029, 0.011, 0.2, 0.2)

$W_L(1)$: (0, 0.2-1*0.029, 0.2-15*0.011, 0, 0)
=>(0, 0.171, 0.035, 0, 0)
=>0+0.171+0.035+0+0 = 0.206

위와 같은 방법으로 List 2는 (F,7,5,T,T), List 3는 (T,2,3,T,T)이다. 따라서 $W_L(2)$=0.455, $W_L(3)$=0.909 이고, WeightedInstanceValue(p) = Max($W_L(n)$), ($W_L(n)$ $W_L(t)$, t = 1, 2, … n) 이 므 로 WeightedInstanceValue(p) = Max(0.206, 0.455, 0.909) = 0.909를 얻을 수 있다.

표 2. Weighted Training Data on the PlayTennis

날씨	온도	습도	바람	Class	가중치
sunny	25	67	weak	no	0.662
sunny	24	70	strong	no	0.768
overcast	25	66	weak	yes	0.673
rainy	21	80	weak	yes	0.635
rainy	18	72	weak	yes	0.636
rainy	19	76	strong	no	0.884
overcast	19	65	strong	yes	0.542
sunny	20	64	weak	no	0.760
sunny	21	60	weak	yes	0.855
rainy	20	76	weak	yes	0.650
sunny	18	61	strong	yes	0.669
overcast	19	69	strong	yes	0.498
overcast	23	65	weak	yes	0.639
rainy	20	80	strong	no	0.909

2.3 예제 결과 분석

예제를 분석하면 다음과 같은 결과를 얻을 수 있다.
1) 기존의 데이터는 각 Instance 간에 같은 Weighted Value 값을 가졌으나 예제에서는 Min=0.498, Max=0.909 으로 Instance 간의 차이를 보인다. 즉 12번 째 데이터는 전문가 의견 과 가장 차이가 큰 데이터 값을 가지고 있고, 14번 째 데이터는 가장 가까운 데이터 값을 가지고 있음을 알 수 있다.
2) 예제에서 사용 된 데이터 속성 값은 명목형(Nominal)과 연속형(Continuous)이 있다. 이중 각 속성값의 Max, Min 값만 갖을 수 있는 명목형 속성 값이 Weighted Value에 영향을 많이 미칠 수 있다. 즉 확연히 구분이 되는 속성 값이 결과에 더 영향을 준다.
3) 전문가 의견을 3가지로 놓고 실험을 하였다. 전문가 의견이 더 많아 질수록 Weighted Value 값은 적은 차이를 갖을 수 있다. 즉 Instance 와 각각의 전문가 의견들과의 차이는 다르나, 결국 Max 함수를 통하여 가장 가까운 값을 선택함으로서 Instance의 중요도를 평가 받는 것이다.

이 결과는 ExpertOpinionList에 따라 각 Instance의 가치 평가를 달리 한다. 따라서 ExpertOpinionList의 선정이 중요하다. 이 List는 전문가와 그와 유사한 신뢰할 수 있는 전문 지식인을 통해서만 작성이 되어야 한다.

위 예제에서의 경우 14번째 데이터가 12번째 데이터보다 약 1.8배에 평가를 받고 있다. 이와 같이 데이터간의 다른 가치 평가를 하는 전문가 의견이 반영된 데이터 전처리는 List에 맞는 특정 Instance를 활용 하거나, List와 상이한 Instance를 잡음 처리하는데 도움을 줄 수 있다.

3. 실험

DOLLS-HI는 홈인테리어 학습 시스템으로 학습 활동 중 학습자 행위 데이터를 추출할 수 있다[4].

본 논문은 이를 이용하여 372명의 학생을 대상으로 행위 데이터를 수집하여 실험하였다.

DOLLS-HI로부터 수집된 학생의 학습 데이터를 기반으로 학생의 학습스타일을 구분해내는 의사결정나무를 제안된 방법으로 생성하였다. Richard Felder와 Linda Silverman[5]가 제안한 학습스타일 중의 하나인 Active/Reflective에 대한 데이터를 이용하였다.

실험 데이터는 Active 186명 Reflective 186명으로 구성되어 있고, 기존의 엔트로피를 이용한 Decision Tree 와 본 논문에서 제안된 엔트로피를 이용한 Decision Tree를 사용하여 각각 3회의 실험을 하였다. 표3 과 표4 실험은 실험 데이터를 372명의 데이터 중 무작위(Random)로 선택하고 그 외 데이터를 Training 데이터로 사용한 실험 한 결과이다. 기존의 방법과 제안한 방법의 Hit Rate 실험 결과가 비슷하다. 그리고 표5 와 표6 실험은 ExpertOpinionList 조건과 유사도가 0.9 이상인 데이터를 Test Data로 사용하고 그 외 데이터를 Training 데이터로 사용하였다. 실험 결과, Hit Rate가 기존에 비해 평균적으로 15.17% 향상된 모습을 보였다. 그리고 표4 와 표6을 보면 두 실험의 Decision Tree의 Node의 수가 모두 감소하고 Height도 모두 낮아져서 기존의 Decision Tree 보다 좋은 성능을 확인 할 수 있다.

표 3. Result of Random Test Data, Hit Rate

	기존 엔트로피	제안한 엔트로피
1	55.00%	57.50%
2	55.00%	52.50%
3	50.00%	47.50%
평균	53.33%	52.50%

표 4. Result of Random Test Data, Improved DT

	Node		Height	
	기존	제안	기존	제안
1	169	123	68	39
2	151	127	57	42
3	161	121	54	36
평균	160.33	123.67	59.67	39

표 5. Result of Test Data(0.9) near ExpertOpinion, Hit Rate

	기존 엔트로피	제안한 엔트로피
1	29.17%	54.17%
2	54.82%	56.60%
3	50.00%	68.75%
평균	44.67%	59.84%

표 6. Result of Test Data(0.9) near ExpertOpinion, Improved DT

	Node		Height	
	기존	제안	기존	제안
1	171	133	60	45
2	151	117	53	41
3	151	125	53	44
평균	157.67	125	55.33	43.33

4. 결론 및 향후 연구

본 연구는 데이터베이스에 있는 데이터들을 데이터 마이닝 시 데이터들이 쓰이는 목적에 따른 예상되는 데이터 속성들을 전문가 의견을 통해 타당성을 평가 하고 전문가 의견에 가까운 Instance를 높은 가중치를 줌으로써 그것을 더 잘 반영하게 하였다.

본 연구결과를 통해서 데이터 Instance들은 데이터 마이닝 시에 목적에 따라 각기 다른 가치 평가를 통해 적용될 수 있음을 알 수 있다. 따라서 목적에 맞는 전문가 의견 구성이 중요 하며 이를 통해 중요한 데이터와 그렇지 않은 데이터를 차별화 함으로써 의사결정자에게 좀 더 의미 있는 데이터를 전달 할 수 있다.

본 연구는 기존의 데이터 마이닝과는 달리 사람의 의견과 지식이 반영되었다는 점이 큰 특징이다. 신뢰되지 않는 데이터에 적합한 사람의 의견이 반영되어 진다면 데이터의 신뢰도를 높이고 이를 의사 결정자에게 제공함으로써 의사 결정시 기존의 것보다 더 나은 전략을 세울 수 있다.

전문가 의견 과 데이터와의 유사도를 통해 가중치를 적용 하므로 유사도를 비교하는 방법이 중요하다. 본 논문이 제안한 방법은 전체 속성을 전부 비교 하였으나, 데이터 Instance에 따라 특성을 나타내는 주요한 속성들이 다를 수 있고, 적은 수의 속성 값이나 각기 다른 속성들로 전문가 의견이 선정된다면 가중치를 구하는 다른 방법이 필요하다. 또한 데이터 속성 값으로 명목형(Nominal)과 연속형(Continuous)을 사용하였다. 이 외에 순서형(Ordinal)을 속성 값으로 갖는 데이터의 비교 방법이 필요하다.

참 고 문 헌

[1] Li-Quan Dong, TaeBok Yoon, Jee-Hyong Lee, Keon-Myong Lee, "Decision Tree Algorithm Improved with a Time-weighted Entropy", SCIS&ISIS 2006, pp.235-239, 2006.

[2] 박우창, 승현우, 용환승, 최기헌, 데이터 마이닝, 자유아카데미, 2003.

[3] 손승현, 김재련, "시간 가중치를 고려한 연관규칙", 산업경영시스템학회논문지, 제 23권 제 61집, pp.453-460, 2000.

[4] Hyun Jin Cha, Yong Se Kim, Seon Hee Park, Tae Bok Yoon, Young Mo Jung, Ji Hyung Lee, "User Interface Behaviors for the Customization of Learning Interfaces in an Intelligent Tutoring System", ITS 2006, pp.513-524, 2006.

[5] Richard Felder, Linda Silverman, "learning and teaching styles in Engineering Education", Engineering Education, pp.674-681, 1988.