시간 가중치 엔트로피를 이용한 결정 트리 생성 알고리즘

동립권[0] 이지형
성균관대학교 정보통신공학부 컴퓨터공학과
dongliquan@skku.edu[0] jhlee@ece.skku.ac.kr

ID3 Algorithm Improved with Time-weighted Entropy

Liquan Dong[0] Jeehyong Lee
Department of Computer Engineering, Sungkyunkwan University

요 약

결정 트리(Decision Tree)는 주어진 데이터의 경향을 학습하는 데 사용되는 대표적인 방식이다. 이것은 주어진 데이터를 구조화하기 위하여 데이터의 속성과 정보의 엔트로피에 기반을 둔 정보획득량을 이용한다. 본 논문에서는 유비쿼터스 환경에서 사용자 프로파일 정보처럼 시간에 따라 그 경향이 변하는 데이터에 유용하게 적용할 수 있는 시간 가중치 엔트로피를 정의한다. 그리고 ID3 알고리즘을 기반으로 새롭게 제안하는 시간 가중치 엔트로피를 이용하는 향상된 ID3 알고리즘을 쓰고 사용자의 경향을 분석한다. 본 논문에서 제안하는 엔트로피를 이용하는 방식은 데이터들의 시간에 관한 영향을 고려해서 기준방식보다 분석결과가 더욱 유리하다. 두 방식의 비교 테스트 결과를 보면 시간 가중치 엔트로피를 이용하는 알고리즘은 기존의 ID3 알고리즘보다 구성된 트리의 구조가 매우 간단하고 유리하다.

1. 서 론

데이터 마이닝(Data Minning)은 수많은 데이터 가운데 숨겨져 있는 유용하게 활용될 수 있는 지식을 효과적으로 찾아내는 지식 탐사의 한 연구 분야이다. 데이터 마이닝의 기법 중에는 연합(Association), 분류(Classification), 집단화(Clustering)등이 있다.

주로 사용하고 있는 분류 방법에는 Bayesian 분류, K Nearest Neighbors, 유전자 알고리즘(Genetic Algorithm), 신경망(Neural Network), Rule-Based 알고리즘, 결정 트리(Decision Tree)등 여러 가지가 있고, 그중에서 결정 트리는 결과가 간단한 트리 형식으로 구성이 되므로 사람들이 쉽게 이해하고 설명할 수 있다는 장점을 갖고 있기 때문에 데이터 마이닝 작업에 많이 쓰고 있다.

결정 트리는 데이터를 구성하는 속성(Attribute)과 클래스(Class)와의 관계를 규명하기 위해 데이터 집합(Data Set)을 부분 집합(Subset)으로 분할하고, 분할된 집합의 특성을 규명하는 데 사용되는 방법이다. 제공하는 결정 트리 알고리즘도 여러 가지가 있다. ID3, ID5R, C4.5, CART(Classification And Regression Trees),SPRINT(Scalable PaRallelizable INduction of decision Trees)등이 있고, 그 중에서 대표적인 것은 1986년에 Quinlan이 제안한 ID3 알고리즘이다[1,2,3,4].

ID3 알고리즘은 결정노드에서 분할 기준이 되는 최적 검사 속성을 선택하기 위해서 속성 선택 척도로 정보획득량(Information Gain)을 사용한다. 이 척도는 정보이론의 엔트로피(Entropy) 개념을 사용하는데 엔트로피 값은 작은 값을 취한다. 따라서 검사 속성 중에서 가장 작은 엔트로피 값을 갖고 있는 속성을 선택하게 된다.

유비쿼터스(Ubiquitous) 환경에서 사용자 프로파일 정보에 관련한 데이터들을 데이터베이스에서 저장해 놓고 사용자에게 다양한 서비스를 제공하기 위해서 사용자 프로파일 정보 데이터들을 학습시키고 사용자의 경향을 분석할 필요성이 있다. 그런데, 실제 유비쿼터스 환경에서 사용자의 경향은 시간에 따라 바뀌는 경우가 많다. 예를 들면, 사용자가 몇 개월 전에 어떤 상황에서 어떤 일을 했었는데 최근에 같은 상황에서 그 일을 안 하거나 다른 상황에서 그 일을 하게 된다. 따라서 사용자 프로파일 데이터들을 분석할 때 과거의 데이터보다는 최근의 데이터에 더 많이 비중을 둘 필요가 있다. 그러나 이와 같이 시간에 따라서 경향이 바뀌는 데이터들을 분석할 때 ID3 알고리즘은 시간요소를 적절히 표현할 수 없다. 그래서 본 논문에서 시간을 고려하여 최근 데이터들의 경향을 더 잘 표현할 수 있도록 시간 가중치 엔트로피를 이용한 결정 트리 생성 알고리즘을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 제안된 시간 가중치 엔트로피 이용한 알고리즘을 소개한다. 3장에서는 ID3 알고리즘과 비교하기 위해서 실험을 하고, 4장에서는 결론 및 향후 연구를 맺는다.

2. 제안하는 알고리즘

우선 ID3 알고리즘에서 정의하는 엔트로피공식을 소개한다. 엔트로피는

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

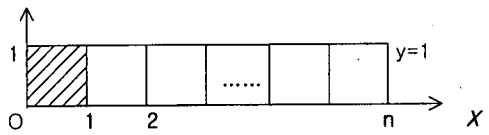
으로 정의한다. S는 사례들의 집합이고, T_c는 사례들이 속하는 클래스의 총 개수이며, P_c는 S중에서 클래스 c에 속하는 사례들의 비율이고, 즉 S에 n개의 사례가 있고 이중 n_c개가 있다면 $P_c = n_c/n$가 된다. 그리고 $0\log_2 0 = 0$로 정의한다.

S중에서 c에 속하는 것이 하나이면 $P_c = 1/n$이고 두 개이면 $2/n$이 되므로 S의 각 사례가 S의 엔트로피에 미치는 영향은 $1/n$에 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 적을 수 있다. $c(i)$는 i번째 사례가 속하는 클래스이다.

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

이를 그림으로 표현하면 <그림1>과 같다.

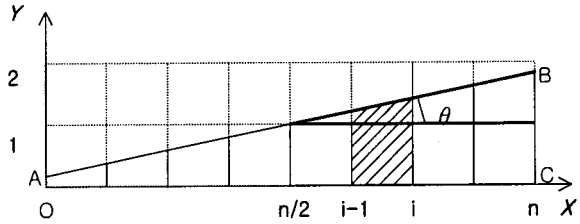


<그림1> 각 사례들에 대한 동등 가중치

그림에서는 각 사례의 가중치는 하나의 사각형 면적으로 표현되어 있다. 예를 들면, 위의 그림 중 회색부분은 첫 번째 사례의 가중치를 나타난다.

ID3 알고리즘은 <그림1>에서 보듯이 모든 데이터의 가중치 값은 모두 1이다. 여기서 n값을 시간이라고 생각한다면, n이 크면 클수록 데이터가 더 최근의 것임을 의미한다. 그러므로 n에 따라서 대응하는 면적도 커져야 한다. 그런데 ID3 알고리즘은 모든 데이터의 가중치 값, 즉 면적이 모두 같기 때문에 시간에 대한 영향을 표현할 수 없다.

문제를 일반적으로 정의하면, 최근 데이터의 영향이 더 커야 한다면 n에 따라서 가중치 값도 커지고, 옛날 데이터의 영향이 더 커야 한다면 n에 따라서 가중치 값이 작아져야 한다. 이 일반화된 정의를 가지고 가중치 값을 결정할 수 있는 그림을 그리면 <그림2>와 같이 나타날 수 있다. 그림에서 보듯이 <그림2>의 OABC의 면적과 <그림1>에서 n개의 정사각형 면적의 합이 같다. 즉, <그림2>에서 n개 사례가 갖는 가중치의 합과 일반적으로 사용되는 엔트로피에서 n개의 사례가 갖는 가중치의 합이 같도록 정의 하였다.



<그림2> 각 사례들에 대한 가중치

새로운 가중치는 직선 AB는 $y=(x-n/2)\tan\theta+1$로 결정된다. <그림2>에서 나타난 그림자의 면적이 바로 i번째 사례의 가중치 값으로 사용하였다. i번째 데이터의 가중치 $W(i)$를 구하면 이래와 같다.

$$W(i) = \frac{2i-n-1}{2}\tan\theta + 1, \quad (-\frac{2}{n} \le \tan\theta \le \frac{2}{n})$$

$\tan\theta$ 범위를 보면 $\tan\theta = 2/n$일 때 n번째 사례의 가중치 값은 가장 크며 최근 데이터 n의 영향이 최대가 되고, $\tan\theta = 0$일 때 가중치 값은 1이 되면서 ID3 알고리즘이 되고, $\tan\theta = -2/n$일 때 n번째 사례의 가중치 값은 가장 작으며 최근 데이터 n의 영향이 최소가 된다.

새로운 가중치를 이용한 엔트로피를 산출하는 공식은 다음과 같다.

$$Entropy\ (S) = -\sum_{c=1}^{T_c} \frac{\sum_{c(i)=c} W(i)}{\sum_{i=1}^{n} W(i)} \log_2 \frac{\sum_{c(i)=c} W(i)}{\sum_{i=1}^{n} W(i)}$$

이 엔트로피는 시간 가중치 엔트로피(Time-weighted Entropy)라고 정의한다. 본 논문에서는 ID3 알고리즘과 시간 가중치 엔트로피를 사용하여 결정 트리를 구성한다.

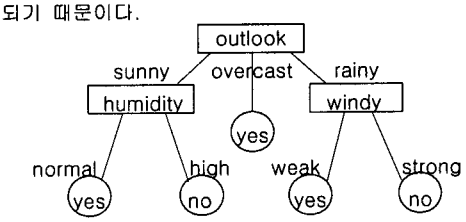
3. 실험

시간 가중치 엔트로피를 사용하는 알고리즘과 기존의 ID3 알고리즘을 비교하기 위해서 참고문헌[5]에서 사용된 테니스를 데이터를 이용하였다. 즉 사용자가 outlook, temperature, humidity, windy 에 따라서 테니스를 했는가 하지 않았는가의 데이터를 결정 트리를 이용하여 학습하는 것이다<표1>.

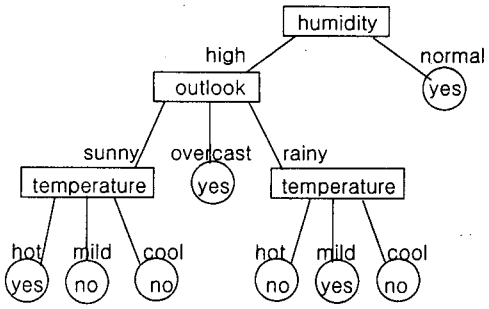
time	outlook	temperature	humidity	windy	class	weight
1	sunny	hot	high	weak	no	0.048
2	sunny	hot	high	strong	no	0.143
3	overcast	hot	high	weak	yes	0.238
4	rainy	mild	high	weak	yes	0.333
5	rainy	cool	normal	weak	yes	0.429
6	rainy	cool	normal	strong	no	0.524
7	overcast	cool	normal	strong	yes	0.619
8	sunny	mild	high	weak	no	0.714
9	sunny	cool	normal	weak	yes	0.810
10	rainy	mild	normal	weak	yes	0.905
11	sunny	mild	normal	strong	yes	1.000
12	overcast	mild	high	strong	yes	1.095
13	overcast	hot	normal	weak	yes	1.190
14	rainy	mild	high	strong	no	1.286
15	overcast	cool	high	weak	yes	1.381
16	rainy	hot	high	strong	no	1.476
17	overcast	mild	normal	weak	yes	1.571
18	sunny	hot	high	weak	yes	1.667
19	sunny	cool	high	strong	no	1.762
20	rainy	hot	normal	weak	yes	1.857
21	rainy	cool	high	strong	no	1.952

<표1> 테니스 데이트

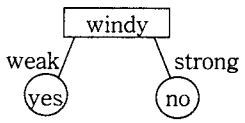
그러나 이러한 사용자 관련 데이터는 시간에 따라 변화하기 때문에 기존의 데이터와는 다른 경향을 갖고 있는 새로운 데이터를 추가하여 실험하였다. <표1>의 1~14는 기존의 데이터이며 15~21 데이터는 새로 추가된 것이다. 그리고 사용자가 테니스를 치는 경향의 변화를 표현하기 위해서 15~21은 기존과는 다른 경향을 갖도록 추가하였다. 즉, 1~14는 <그림3>과 같은 경향은 보이지만 8~21은 <그림4> 와 같이 되도록 15~21 데이터를 겹치도록 한 것은 사용자의 경향은 서서히 바뀔 것으로 예상되기 때문이다.



<그림3> 1~14번 데이터 ID3 알고리즘으로 생성된 결정 트리

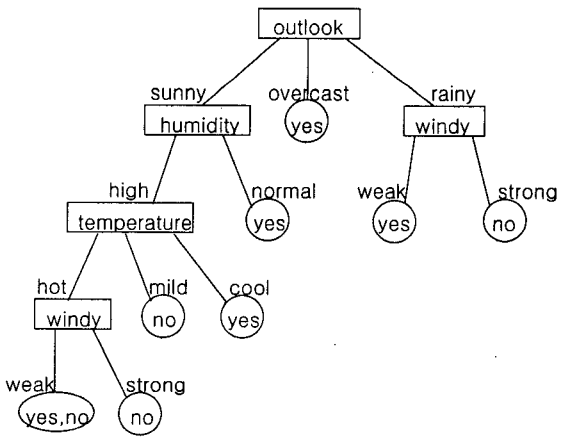


<그림4> 8~21번 데이터 ID3 알고리즘으로 생성된 결정 트리



<그림5> 15~21번 데이터 ID3 알고리즘으로 생성된 결정 트리

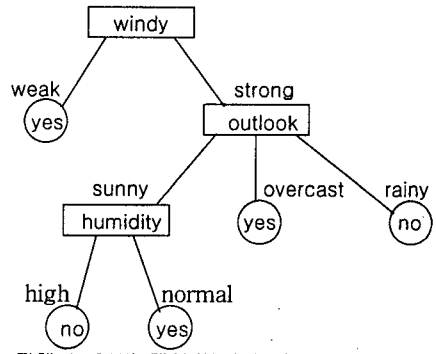
이렇게 변화하는 사용자 경향을 적절히 학습하기 위해서는 이미 다른 경향을 보이는 과거 데이터보다는 현재 데이터에 관심을 갖아야 하는데 어려운 점은 과연 어느 데이터부터가 새로운 경향을 보이는가를 어떻게 알아낼 수 있는가이다. 예를 들어 데이터 8~21이 새로운 경향을 보인다고 생각하여 여기에 ID3 알고리즘을 적용하면 바뀐 경향을 잘 반영하는 결정 트리를 얻을 수 있지만, 모든 데이터가 동일하게 중요하다는 판단에 1~21까지의 데이터에 ID3 알고리즘을 적용하여 결정트리를 구하면 <그림6>과 같이 복잡한 트리를 얻게 된다. 그 이유를 이미 지난 간 과거의 데이터와 새로운 경향의 데이터가 서로 섞여 있기 때문이다. 또한 최신 데이터가 새로운 경향을 반영한다고 하여 지나치게 최신 데이터만을 고려해도 적절한 결절 트리를 얻을 수 없게 된다. 예를 들어 데이터 15~21에 대한 결정 트리를 구하면 <그림5>와 같이 지나치게 간단한 것은 얻게 된다. 이와 같이 변화하는 데이터에서 경향을 반영하는 결정 트리를 생성하기는 쉽지 않다.



<그림6> 1~21번 데이터 ID3 알고리즘으로 생성된 결정 트리

이를 위해서 본 논문에서는 시간 가중치 엔트로피와 ID3 알고리즘을 이용하여 결정 트리를 생성하는 방법을 제안한다. 즉, ID3 알고리즘으로 결정 트리를 생성하는 과정에서 기존의 엔트로피 대신 새로 제안하는 엔트로피를 사용한다. 그리고 트리 생성 중 특정 노드의 엔트로피가 사전에 정한 기존 값보다 작으면 그 노드에 속하는 데이터가 가장 많이 속하는 클래스로 대치하는 방법을 적용하였다.

이러한 방법을 위의 데이터에 적용해보면 <그림7>과 같은 결정 트리가 나온다. 최근 데이터에 중점 둔 결과 상위속성으로 windy가 선택되었다. 그리고 windy = weak 인 경우를 살펴보면 과거의 두 경우(1,8)에만 no 이고 최신 데이터를 포함한 나머지는 yes이다. 이 경우의 엔트로피로 구하면 과거 데이터의 가중치가 낮은 관계로 0.095가 나와 기준 값 0.1에 미달하므로 가장 많이 나타는 클래스인 yes로 치환하였다.



<그림7> 전체 1~21번 데이터 사간 가중치 엔트로피 알고리즘으로

4. 결론 및 향후연구

본 논문에서 제안한 시간 가중치 엔트로피공식을 사용하여 구성 된 트리 구조가 원래의 ID3 알고리즘으로 구성된 트리보다 더 간단하고 유리하다. 그 이유는 ID3 알고리즘은 최근 경향을 찾기 어려운 것과 달리 논문에서 제안한 구조는 최근 데이터에 더 중점을 두고 있기 때문에 전체 데이터를 고려해도 최근 데이터의 영향을 볼 수 있다.

그래서 유비쿼터스 환경에서 사용자 프로파일 정보가 시간에 따라 계속 바꾸어도 그 경향이 변하는 데이터에 유용하게 적용할 수 있는 시간 가중치 엔트로피를 사용하면 사용자의 경향을 잘 분석 할 수 있다. 분석결과를 가지고 사용자에게 다양한 서비스를 제공할 수 있다.

향후연구를 데이터 량이 크면 클수록 트리의 Pruning방법하고 새로운 데이터를 추가할 수 있는 incremental방법도 더 추가해서 연구할 예정이다.

5. 참고문헌

[1] Paul Utogff, "Incremental induction of decision trees," Machine Learning, 4(2), 161-186, 1989.
[2] Margaret H. Dunham, Data Mining Introductory and Advanced Topics, Prentice Hall, 2002.
[3] Mehmed Kantardzic, Data Mining Concepts, Models, Methods, and Algorithms, Wiley-IEEE Press, 2002.
[4] John Shafer, Rakesh Agrawal, Manish Mehta, "SPRINT: A Scalable Parallel Classifier for Data Mining,"Research report, IBM Almaden Research Center, San Jose, California, 1996.
[5] Tom M. Mitchell, Machine Learning, McGraw-Hill Science/Engineering/Math, 1997.