데이터의 다중 추상화 수준을 위한 결정 트리*

정민아○ 이도헌
전남대학교 컴퓨터정보학부
{majung, dhlee}@dblab.chonnam.ac.kr

Decision Trees For Multiple Abstraction Level of Data

Mina Jeong○ Doheon Lee
School of Computer and Information, Chonnam National University

요 약

데이터 분류(classification)란 이미 분류된 객체 집단군 즉, 학습 데이터에 대한 분석을 바탕으로 아직 분류되지 않는 객체의 소속 집단을 결정하는 작업이다. 현재까지 제안된 여러 가지 분류 모델 중 결정 트리(decision tree)는 인간이 이해하기 쉬운 형태를 갖고 있기 때문에 탐사적인 데이터 마이닝(exploatory)작업에 특히 유용하다. 본 논문에서는 결정 트리 분류에 다중 추상화 수준 문제(multiple abstraction level problem)를 소개하고 이러한 문제를 다루기 위한 실용적인 방법을 제안한다. 데이터의 다중 추상화 수준 문제를 해결하기 위해 추상화 수준을 강제로 같게 하는 것이 문제를 해결할 수 없다는 것을 보인 후, 데이터 값들 사이의 일반화, 세분화 관련성을 그대로 유지하면서 존재하는 정보를 유용화할 수 있는 방법을 제시한다.

1. 서론

데이터 분류(classification)란 이미 분류된 객체 집단군 즉, 학습 데이타에 대한 분석을 바탕으로 아직 분류되지 않는 객체의 소속 집단을 결정하는 작업이다. 현재까지 제안된 여러 가지 분류 모델 중 결정 트리(decision tree)는 인간이 이해하기 쉬운 형태를 갖고 있기 때문에 탐사적인 데이터 마이닝(exploratory data mining) 작업에 특히 유용하다 [1][2][3].

기존의 데이터 분류 기법은 대부분 단일 정보원(information source)으로부터 미리 잘 정리된 학습 데이터를 획득하는 것을 가정하고 있다. 하지만, 실제적인 데이터 마이닝 환경에서는 여러 데이터베이스 혹은 파일 시스템으로부터 학습 데이터를 추출해야 하는 경우가 일반적이다. 이 때, 각 정보원들은 각자의 운용 목적에 따라 독자적으로 구축, 관리되어 온 것이기 때문에, 추출한 데이터는 서로 다른 추상화 수준으로 표현되어 있다. 예를 들어, 네트워크 알람 관리 시스템에서 일부 지역 네트워크 알람 관리자는 '라우터 비정상'과 같이 높은 추상화 수준으로 알람을 보고하지만, 다른 지역 네트워크 알람 관리자는 '유형-32 라우터 메시지 오버플로우'와 같이 낮은 추상화 수준으로 알람을 보고할 수 있다 [4].

이러한 데이터 품질(data quality) 문제를 해결하기 위해 다양한 데이터 정화(data cleansing) 도구가 개발되고 있다 [5][6][7]. 데이터 정화 도구를 이용하여 다중 추상화 수준을 해결하는 시도로서 두 가지 방안을 고려할 수 있다. 첫 번째 방안은 상향 평준화 방법으로서, 낮은 추상화 수준을 가진 데이터를 일정한 수준으로 높이는 방법이다. 하지만, 데이터가 높은 추상화 수준을 가질수록 구체성이 떨어지기 때문에 결국 얻게 되는 분류 모델의 구체성이 떨어지게 된다. 결국 이미 확보한 유용한 정보까지 활용하지 못하는 정보 손실 문제가 발생한다. 두 번째 방안은 하향 평준화 방법으로서, 높은 추상화 수준을 가진 데이터를 일정한 수준으로 낮추는 방법이다. 하지만 데이터의 추상화를 낮추기 위해서는 부가적인 정보가 필요하다. 예를 들어 확보한 데이터는 '콜라'인데, '펩시 콜라'인지 '코카 콜라'인지 알기 위해서는 별도의 정보 획득이 필요하다. 따라서 이 방안 역시 실제 환경에서 적용하기 힘들다. 결국 데이터 정화 도구를 이용한 방법으로는 다중 추상화 수준 문제를 해결하기 곤란하다.

본 논문에서는 결정 트리 분류에 다중 추상화 수준 문제(multiple abstraction level problem)를 소개하고 이러한 문제를 다루기 위한 실용적인 방법을 제안한다. 데이터 값을 강제로 일반화하거나 세분화하기 보다는 존재하는 정보를 그대로 사용하며, 데이터 값들 사이의 일반화, 세분화 관련성은 그대로 유지되도록 한다.

본 논문의 구성은 다음과 같다. 2장에서는 결정 트리를 구축하는 기존의 알고리즘과 학습 데이터가 다중 추상화 수준으로 표현될 때 결정 트리를 구축하는 방법을 기술한다. 3장에서는 제안된 방법을 적용한 후 얻어진 결정 트리를 이용해 분류되지 않는 객체에 클래스를 할당하는 방법을 제시한다. 4장에서는 본 논문에 대한 결론 및 향후 연구 방향을 제시한다.

2. 결정 트리 구축

2.1 결정 트리 구축을 위한 알고리즘

결정 트리는 학습 데이터의 순환적 분할을 통해 구축된다.

* 본 연구는 첨단정보기술연구센터(AITrc)를 통해서 한국과학재단으로부터 지원을 받았음.

결정 트리를 구축하는 첫 번째 단계는 학습 데이터의 최적 분할 속성의 선택이다. 다음 단계에서 학습 데이터는 그 속성값에 따라 부분집합들로 분할된다. 이러한 선택-분할 과정은 각 부분집합에 순환적으로 적용된다. 그림 1은 결정 트리를 구축하기 위한 알고리즘이다.

```
ConstructTree(Node ThisNode, Relation R)
{
(1) Attr = SelectBestSplit(R);
    //Select the best-split attribute of R
(2) ThisNode.label = Attr;
   //Mark the label of ThisNode as Attr
(3) For each attribute value x of Attr
(4)  {
        // Select records whose Attr attributes are 'X'
(5)      NewR = SelectRecords(R, Attr, x);
        // Make a new decision node
(6)      NewNode = NewNode();
        // Make NewNode a child node of ThisNode
(7)      NewNode.parent = ThisNode;
        // call this procedure itself for the child node
(8)     ConstructTree(NewNode, NewR);
    }
}
```

그림 1. 결정 트리를 구축하기 위한 알고리즘

학습 데이터에 다중 추상화 수준이 나타나는 레코드들이 포함되어 있을 때 최적 분할 속성의 선택과 최적 분할 속성에 따라 학습 데이터를 분할하는 부분은 다시 수정되어야 한다.

2.1 다중 추상화 수준을 나타내는 학습 데이터의 분할

학습 데이터의 분할 부분이 수정되어야 하는 이유를 예제를 통해 제시한다. 표 1과 같이 고객 정보에 대한 학습 데이터를 가지고 있다고 가정한다.

RID	RA	Gender	Income	FVB
t1	East	Male	90k	E2
t2	East	Male	70k	E1
t3	Far-East	Female	80k	E1
t4	Mid-East	Male	50k	E2
t5	Mid-East	Female	30k	E2
t6	West	Male	90k	W1
t7	West	Male	50k	W1
t8	West	Female	100k	W2
t9	West	Female	40k	W2
t10	West	Female	50k	W2

표 1 고객 정보에 대한 학습 데이터

속성 'RA'가 클래스 'FVB'에 대하여 최적 분할 속성으로 선택되었다고 가정한다. 최적 분할 속성을 선택하는 방법은 2.2절에서 자세히 언급한다. 먼저 'RA'값에 따라 학습 데이터를 분할한다. 기존의 분할 알고리즘은 {t1,t2}, {t3}, {t4, t5}, {t6, t7, t8, t9, t10}으로 분할을 한다. 그러나 그러한 분할은 'Far-East'와 'Mid-East'가 'East'에 속한다는 사실을 나타내지 않는다. 그러므로 첫 번째 집합에 포함시킬 수 있고, 결과는 {t1,t2,t3,t4,t5}, {t3}, {t4, t5}, {t6,t7,t8,t9,t10}이 된다. 그러나 첫 번째 고객의 'RA'값이 'East'로 기록되었더라도 고객의 'RA'값이 'Far-East'거나 'Mid-East'가 될 수 있다는 사실을 나타내고 있지 않다.

학습 데이터는 전체 영역의 값에 대한 분포를 반영하기 때문에 학습 데이터의 분포로부터 실제 'RA'값의 확률을 얻을 수 있다. 세 개의 레코드 t3, t4, t5는 'East'의 'RA' 값으로 세분화되어 졌다. 그들 중 한 레코드 t3는 'Far-East'의 값을 갖고 있으며 다른 두 레코드 t4와 t5는 'Mid-East'의 값을 갖고 있다. 이러한 분포에서 t1(또는 t2)의 'RA'값이 실제 'Far-East'일 확률은 1/3=33%이다. 비슷하게 t1(또는 t2)의 'RA'값이 'Mid-East'는 2/3=67%인 확률을 갖는다. 결과적으로 학습 데이터는 {t1,t2,t3,t4,t5},{t1/0.33, t2/0.33, t3}, {t1/0.67, t2/0.67, t4,t5}, {t6,t7,t8,t9,t10}로 분할이 되며, ti/μ의 ti는 소속 정도 μ 값으로 집합에 속한다는 것을 나타낸다. 정의 1과 2는 이러한 사실을 형식화한 것이다. 제시한 예제에서와 같이 집합내의 레코드에 대한 부분적 소속 정도를 다루는 것이 필요하다. 본 논문에서는 이러한 소속 정도를 표현하기 위해 퍼지 개념을 도입하였다.

정의 1 (퍼지 관계)
퍼지 관계 T는 다음과 같이 정의한다.

$$T = \{(t, \mu_T(t)) \mid t \text{ 는 보통의 레코드},\ \mu_T(t)\text{는 T내의 t의 소속 정도}\}$$

소속 정도 $\mu_T(t)$는 레코드가 얼마나 완전하게 집합에 속하는가를 표현하기 위해 각 레코드에 첨가된다.만약 $\mu_T(t)=1$이면 완전한 소속 정도를 의미하고 $\mu_T(t)<1$은 부분적 소속 정도를 의미한다.

정의 2 (학습 데이터 분할)
퍼지 관계 T(학습 데이터)와 속성 X의 영역에 대한 ISA 계층구조 H가 주어졌을 때 'X가 x'와 같은 조건을 갖는 T로부터의 선택인 SS(T, H, X, x)를 다음과 같이 정의한다.

$$SS(T,H,X,x) = SS_{direct}(T,X,x) \cup SS_{descendent}(T,H,X,x) \cup SS_{antecedent}(T,H,X,x),$$

where

$$SS_{direct}(T,X,x) = \{(t, \mu(t)) \mid t \in T,\ t.X = x,\ \mu(t) = \mu_T(t)\},$$

$$SS_{descendent}(T,H,X,x) = \{(t, \mu(t)) \mid t \in T,\ t.X \in DESC(x,H),\ \mu(t) = \mu_T(t)\}, \text{and}$$

$$SS_{antecedent}(T,H,X,x) = \{(t, \mu(t)) \mid t \in T,\ t.X \in ANTE(x,H),\ \mu(t) = \mu_T(t) \times (Card(\{s \mid s \in T,\ s.X = x \text{ or } s.X \in DESC(x,H)\}) / Card(\{s \mid s \in T,\ s.X \in DESC(t.X,H)\}))\}.$$

(1)

ANTE(x)와 DESC(x)는 ISA 계층구조에서 조상과 후손으로써 나타나는 값들의 집합을 의미하며, Card(T) = $\Sigma\,\mu_T(t)$이다.

X값이 x인 레코드들을 집합 T로부터 선택한 후의 결과 집합 즉 SS(T,H,X,x)는 세 부분으로 구성된다. 첫 번째 부분집합 $SS_{direct}(T,X,x)$는 X값이 x에 글자 그대로 맞는 집합이다. 이 때 $\mu(t)$값은 집합 T에서 t의 소속 정도를 나타내는 $\mu_T(t)$이다. 두 번째 부분집합 $SS_{descendent}(T,H,X,x)$는 X값이 x의 세분화인 레코드들의 집합이다. 앞 예제에서 'Fat-East'는 'East'의 세분화를 나타낸다. 이러한 경우에 $\mu(t)$값은 $\mu_T(t)$인데 세분화 값은 일반화 값에 완전히 속한다. 세 번째 부분집합 $SS_{antecedent}(T,H,X,x)$는 X값이 x의 일반화인 레코드들의 집합이다. 이러한 경우 $\mu(t)$값은 1.0이하로 부분적 소속 정도를 갖는다. 이것은 일반화 값이 세분화 값에 꼭 속하는 것은 아니기 때문이다.

2.2 최적 분할 속성을 선택하는 기준치 확장

최적 분할 속성은 학습 데이터를 가장 동일한 부분집합들로 분할하는 속성이다. 현재까지 제안된 몇 가지 기준치가 집합의 이질성을 평가하기 위해 존재한다. 본 논문에서는 정보 이론의

기준치를 적용하는데, 이것은 실제 데이터 마이닝 시스템에서 가장 많이 사용되는 것 중의 하나이다. 퍼지 관계로 학습 데이터를 표현하기 때문에 엔트로피의 기준치를 확장함으로써 퍼지 테이블의 이질성을 측정하는 방법을 정의한다.

정의 3.(퍼지 테이블의 엔트로피)
주어진 퍼지 테이블 T가 구별되는 속성 C에 따라 T^c1,..., T^{ck}k로 분할된다고 가정한다. 이때, T^cj= {(t, μ(t)) | t ∈ T, t.C = cj, μ(t)= μ_T(t)}이다. C에 대한 T의 엔트로피는 $info^C(T)$로 표현하고 다음과 같이 정의한다.

$$info^C(T) = -\sum_{j=1,\dots,k} [Card(T^Cj)/Card(T) \times \log_2(Card(T^Cj)/Card(T))],$$

$$where\ Card(T^Cj) = \sum_{t \in T^Cj} \mu_{T^Cj}(t) 이고\ Card(T) = \sum_{t \in T} \mu_T(t)$$

3. 결정 트리에 의한 클래스 할당

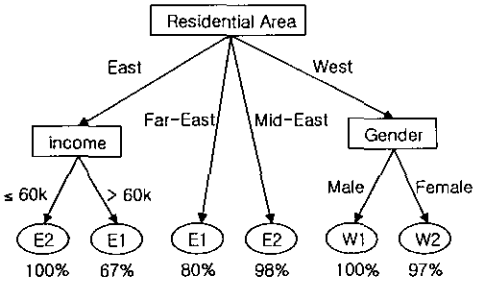

그림 2. 신뢰도를 갖는 결정 트리의 예

제안된 방법을 적용한 후 그림 2에 보여진 결정 트리를 얻었다고 가정한다. 각 단말 노드의 레이블은 결정을 하기 위한 신뢰도를 표현한다. (East, Male, 85k, Unknown)과 같은 목표 레코드를 가지고 있다고 하자. 마지막 속성 'FVB' 즉 클래스 레이블이 알려지지 않았다고 할 때 레코드의 'FVB' 속성을 결정한다고 하자. 기존의 방법들은 뿌리 노드로부터 첫 번째 가지와 'income' 노드의 두 번째 가지를 따라 67%의 신뢰도를 갖는 'E1'에 할당된다. 그러나 'East'는 사실 'Mid-East'일수 있다는 사실을 설명해야 한다. 'RA' 속성이 'East'의 세분화인 'Mid-East'값을 갖는 학습 데이터 중 85% 레코드가 있다고 가정한다. 이러한 분포를 고려하여 뿌리 노드의 세 번째 가지를 따를 수 있으며 'E2'에 신뢰도 85% × 98%=83%로 할당된다. 'E2'는 83%의 신뢰도를 갖는 반면 'E1'은 67%의 신뢰도를 갖는다. 이에 따라 'E2'가 83%의 신뢰도로 할당된다. 그림 3의 알고리즘은 할당 과정을 표현한다.

```
AssignClass(DecisionNode Attr, Record R)
{
(1)  If Attr is a terminal node, return (Attr.Decision);
(2)   Child = the node followed by the branch with a label
              identical to R.Attr.
(3)  Answer = ConsultTree(Child, R);
(4)  For each branch with a label
              that is generalization of R.Attr
     {
(5)       Child = the node followed by this branch;
(6)       Temp = ConsultTree(Child, R);
(7)       If Temp.Confidence > Answer.Confidence,
                    then Answer = Temp;
     }
(8)  For each branch with a label
              that is a specialization of R.Attr
     {
(9)       Child = the node followed by this branch;
(10)      Weight = the ratio of this specialization;
(11)      Temp = Weight × ConsultTree(Child, R);
(12)      If Temp.Confidence > Answer.Confidence,
                    then Answer = Temp;
     }
(13) return(Answer);
}
```

그림 3. 다중 추상화 수준을 갖는 데이터 결정 알고리즘

기존 할당 처리과정은 Line(1)부터 (3)에서 처리되며 데이터의 일반화와 세분화를 고려하여 line(4)에서 (7)까지와 line(8)에서 (12)까지의 루프가 첨가되었다.

4. 결론

본 논문은 결정 트리 분류에 다중 추상화 수준 문제를 도입하였고 이 문제를 다루기 위한 방법을 제안하였다. 데이터 정화 도구를 이용하여 강제로 추상화 수준을 같게 하는 것은 이러한 문제를 해결할 수 없다. 본 논문에서 제안한 방법은 현재 존재하는 정보를 이용하며 데이터 값들 사이의 일반화/세분화 관련성을 고려한다. 세분화된 값이 그 값의 일반화 된 값과 독립적으로 존재할 수 있는 반면, 일반화된 값은 세분화된 값이 존재하는 분포만큼 각각의 세분화된 값과 같이 존재한다. 이러한 부분적 존재 양립성을 표현하기 위해 퍼지 관계를 도입하였다.

현재 제안된 방법들을 구현하고 있으며, 향후 연구로는 실제 데이터를 가지고 성능평가를 하고자 한다.

참고 문헌

[1] J. Gehrke, V. Ganti, R. Ramakrishnan, and W.-Y Loh, "BOAT Optimistic Decision Tree Construction," *In Proc. of ACM SIGMOD Conf.*, Philadelphia, Pennsylvania, June 1999, pp. 169-180

[2] M. Berry and G. Linoff, *Data Mining Techniques For Marketing, Sales, and Customer Support*, Wiley and Sons, 1997

[3] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Pub., 1993

[4] K. Hatonen, M. Klemettinen, H. Mannila, P. Ronkainen, and H. Toivonen, "Knowledge Discovery from Telecommunication Network Alarm Databases," *In Proc. of the 12th International Conference on Data Engineering*, New Orleans, Louisiana, February 1996, pp. 115-122,

[5] R. Wang, V. Storey and C. Firth, "A Framework for Analysis of Data Quality Research," *IEEE Transactions on Knowledge and Engineering*, 7(4), Aug. 1995, pp. 623-640

[6] Trillium Software System, "A Practical Guide to Achieving Enterprise Data Quality," White Paper, Trillium Software, 1998.

[7] Vality Technology Inc., "The Five Legacy Data Contaminants You Will Encounter in Your Warehouse Migration," White Paper, Vality Technology Inc., 1998