属性の数量概念を考慮した例からの学習

1C−2

藤井雅文 中西正和
慶應義塾大学理工学研究科計算機科学専攻

1. はじめに

例からの学習を行うシステムに ID3[1] がある。これは、属性と属性値の対からなる集合によって表される訓練事例を、トップダウン的な分割統治によって分類するシステムである。分類結果は、決定木で表される。分類の仕方は、ある属性を選択し、その属性の属性値について訓練事例を部分集合群に分割する。分割してできたそれぞれの部分集合に対して、同様の分割を繰り返す。これを、部分集合の要素がすべて同じクラスになるまで繰り返す。クラスとは、訓練事例の属している概念のことであるが、例からの学習では訓練事例のクラスがあらかじめわかっている。

このようにして作られる決定木の質は、どの属性について分割をおこなうかということに依存する。ID3 は分割をおこなう属性として、訓練事例の情報量エントロピを計算しその値が小さいものを選択する。これによって、決定木の大きさが最も小さいものが分類結果となる。

ID3 は、分割をおこなう属性として単一の属性を選択する。よって、属性間の関係を考慮したような決定木をつくることはできない。本稿では、属性間に関係がある事例としてマルチプレクサ問題をとりあげ、これを分類する方法について考える。

2. マルチプレクサ問題

マルチプレクサ問題とは、属性値それぞれを数値で表しそれぞれの算術演算を通して得られた数値結果を訓練事例のクラスとしたときに、このような事例を分類する問題である。

例として表1の訓練事例を考える。A、B、Cは、{0、1} を属性値とする属性であり、A、B、C の和を2進数で表したものを事例の属するクラスとしている。

表 1:訓練事例 1

Example	A	B	C	Class
#1	1	0	1	10
#2	0	0	0	00
#3	0	1	0	01
#4	1	1	1	11
#5	0	1	1	10

前述のように、ID3 は分割をおこなう属性として単一の属性を選択するので、この訓練事例のように属性間に関係があるものはうまく学習することができない。

そこで、数量概念を扱うことができる線形論理をこの問題に適用することを考える。

3. 線形論理

線形論理 (linear logic)[2] は、Jean-Yves Girard によって提唱された比較的新しい論理であり、その形式体系を LL と呼ぶ。

LL には、and、or それぞれに乗法的、加法的なものがあり、resource-concious な論理、すなわち、論理概念の中に数量概念が入ったものと考えることができる。

$X \otimes Y \multimap Z$の意味は『XとYをちょうど1回ずつ用いてZを結論することができる』となる。

これを、マルチプレクサ問題に適用してみる。表1の訓練事例について$A \otimes B \otimes C$を考えると、{A、B、C}={0、0、0} ならば、$A \otimes B \otimes C$は0となり、{A、B、C}={1、0、1} ならば、$A \otimes B \otimes C$は2となる。よって、$A \otimes B \otimes C$は {0、1、2、3} の値をとることになる。

4. 線形論理を考慮した分類

$X \otimes Y \multimap Z$を考慮すると表1の訓練事例では、A、B、C、$A \otimes B$、$B \otimes C$、$C \otimes A$、$A \otimes B \otimes C$が分類を

---

learning from examples considering quantitative concept of attributes
Masafumi FUJII, Masakazu NAKANISHI
Department of Computer Science, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa Pref., 223, Japan

おこなう属性の候補となる。一般には n 個の属性があれば 2^n-1 個の候補について情報量エントロピを計算することになる。

このような候補にたいして情報量エントロピを計算するためには、属性値が数値となっている必要がある。表2のような訓練事例が与えられた場合は、Outlook という属性にたいしては sunny=0、overcast=1、rain=2 とし、Temp. にたいしては hot=0、mild=1、cool=2 のようにする。Humid.、Windy にたいしても、同様に属性値に数値を割り当てる。

表 2:訓練事例 2

	Outlook	Temp.	Humid.	Windy	Class
#1	sunny	hot	high	false	N
#2	sunny	hot	high	true	N
#3	overcast	hot	high	false	P
#4	rain	mild	high	false	P
#5	rain	cool	normal	false	P
#6	rain	cool	normal	true	N
#7	overcast	cool	normal	true	P
#8	sunny	mild	high	false	N
#9	sunny	cool	normal	false	P
#10	rain	mild	normal	false	P
#11	sunny	mild	normal	true	P
#12	overcast	mild	high	true	P
#13	overcast	hot	normal	false	P
#14	rain	mild	high	true	N

属性値に数値を割り当てて⊗をもちいた分類を考えてみると、例えば表2で Outlook⊗Temp は 0 から 4 の値をとり、訓練事例が 5 つの部分集合に分割されることになる。Outlook⊗Temp⊗Humid ではさらに多くの部分集合に分割される。情報量エントロピはその性質上、部分集合の数が多くなればなるほど小さい値をとりやすくなる。よって、属性間に⊗の関係が成り立たなくても、その候補に関して分割をおこなってしまう。

この対策として、⊗を含む候補を選択するのは、情報量エントロピの値が 0 となる時、すなわちその候補によって全ての事例が 1 つのクラスからなる部分集合に分割された時のみとする。

5. 実験とその結果

前節で述べた方法で表 1、表 2 の訓練事例を分類した結果を図 1、図 2 に示す。

図 1 をもちいて表 1 の訓練事例にはでてこない $\{A、B、C\}=\{0、0、1\}$、$\{1、0、0\}$、$\{1、1、0\}$ を考えると、たしかに分類されるべきクラスに分類される。よって、表 1 については学習できたことになる。

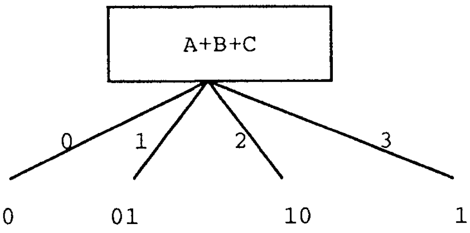

図 1:表 1 の決定木

図 2 は、オリジナルの ID3 に表 2 の訓練事例を与えた時にできる決定木と同じものである。よって、表 2 についても学習をおこなえたことになる。

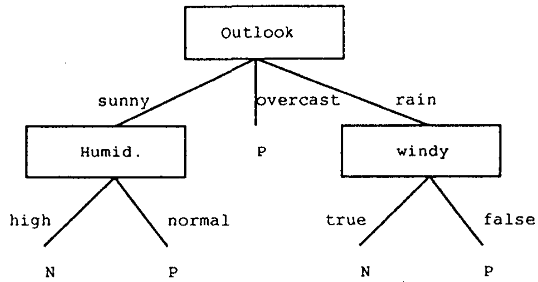

図 2:表 2 の決定木

6. おわりに

今回は、マルチプレクサ問題に対応するために $X \otimes Y \multimap Z$ で表される乗法的な and について考えたが、この他にも線形論理には古典論理にはない乗法的な or がある。これは℘で表され、$X ℘ Y$ の意味は『もし X と Y との一方が成立しなければもう一方が成立する』となる。今後は、乗法的な or をもちいた分類をおこなう属性についても考えていきたい。このことによって線形論理で表すことができる事例は、ID3 でも扱えることができるようになると思われる。

また、今回の方法では、⊗を含む候補を選択する場合を制限しているため、決定木の節が⊗を含むのはその節を展開したものがすべて葉となる時に限られる。つまり、属性間に 2 種類以上の⊗の関係が成り立っている事例は学習できない。このような事例に対応するのも今後の課題である。

参考文献

[1] Quinlan,J.R.(1985). Induction of Decision Trees. *Machine Learning Vol.1* (pp 81-106).

[2] Jean-Yves Girald (1987). Linear Logic. *Theoretical Computer Science 50* (pp 1-102).