決定木構成法における連続値分割について

3J-9

――相互情報量に基づく区間分割――

泉 寛幸　　佐藤 秀樹
(㈱富士通研究所)

1．はじめに

エキスパートシステム開発において，知識ベース構築に多大なコストを要することが問題となっている．このため，従来人手により行われていたこの作業を計算機により支援あるいは自動化するための知識獲得・学習の研究が必要となってきている．故障診断や医療診断のような分類問題においては，訓練事例集合をもとに分類規則を決定木の形式で構成する計算機プログラムの研究開発が行われている[1,2]．ここで，訓練事例は，＜属性，値＞リストと対応する分類クラスとから成る．値は，連続値（例えば，3.298..），離散値（例えば，5），記号の有限集合の要素（例えば，高／中／低）をとる．

決定木構成法では，属性値の領域が連続値の場合に，その領域は有限個の排他的な区間に分割されて取り扱われる．この分割は作成される決定木の性能に影響を与えるので，どのように分割するかが問題となる．本報告では，クラスと区間との間の関係の強さを示す量として閾値情報量を定義し，この閾値情報量が極大になるように領域を区間に分割するという方法を提案する．また，この基準のもとで，訓練事例集合が一様分布や正規分布をなす場合に，閾値がどのように設定されるかについて述べる．

2．定義

我々は，事例の集合を二つのクラスC_1,C_2 に分類する決定木を構成する場合について考える．$C=\{C_1,C_2\}$ をクラス集合と呼ぶ．訓練事例の集合をS_0，S_0の要素数をn_0，S_0内でクラス C_i に対応する事例の集合を S_i,その要素数を n_i とする($i=1,2$; $S_0=S_1\cup S_2$; $n_0=n_1+n_2$)．

訓練事例の＜属性，値＞リストのすべての属性の集合をATとする．属性$A\in AT$は，訓練事例$e\in S$ を引数として，値を出力する関数A(e)とみなす．値として連続値をとる属性を連続値属性と呼ぶ．

S_j($j=0,1,2$) において連続値属性Aの最小値をminAj，最大値をmaxAj とする．値の範囲を閉区間 [minAj, maxAj] とする．ここで，a,b,x を連続値として，$\{x \mid a\leqq x\leqq b\}$ を閉区間［a，b］，$\{x \mid a\leqq x<b\}$ を区間［a，b）と略記する．有限個の区間が共通な要素を持たないとき，排他的な区間とよぶ．

連続値属性Aの値のS_jに対する平均値と分散とをそれぞれ x_j, sj^2とする．$x_1\leqq x_2$と仮定する．

3．決定木構成法

従来の決定木構成法では，以下の手順により決定木を作成する[1,2]

step1.訓練事例集合を木の根ノードに対応づける．

step2.ノードに対応づけられた事例集合が停止条件を満たすならば，その事例をもとにクラスを決定してノードにラベル付けて終了する．

step3.停止条件を満たさなければ，属性集合ATから一つ属性を選択する．

step4.その属性の各値に従って，事例集合を有限個の排他的な部分集合に分割し，各部分集合に対して子ノードを作成して対応づける．もとのノードと各子ノードとの間の枝に属性と値とをラベル付ける．

step5.各子ノードに対してstep2.から繰り返す，

4．連続値の分割の問題

step4.において，その属性が離散値または有限集合の要素を値とする場合には，その各値に従って訓練事例集合を部分集合に分割すれば良い．

一方，属性の領域が連続値の場合には，その領域をあらかじめ有限個の排他的区間に分割しておき，各事例の属性値がどの区間に含まれているかに従って，訓練事例集合を部分集合に分割することが必要となる[2]．

区間の分割の仕方が異なれば，構成される決定木も異なる．決定木が異なれば，その分類能力も異なってくる．従って，分類能力の高い決定木を構成するための効果的な連続値の区間分割を行う閾値決定法の開発が重要となってくる．このため，次節では，閾値情報量の概念を用いる連続値の分割の方法を提案する．

5．閾値情報量

従来，step3.においては，事例集合の分割のための属性として，属性集合ATの要素である属性AとクラスC1とC2との間の相互情報量を計算し，その値が最大のものを選択している[1,2]．この相互情報量は，属性Aの値を知ったときに，対応する事例がクラスC_1かC_2に属する可能性に関する情報を表す量である．この相互情報量に基づく属性選択法は，比較的分類能力の高い決定木が得られると言われている[1]．同様に，我々も属性選択のため，この相互情報量を利用している．

さらに，この相互情報量の概念が，連続値の分割のための閾値を決定するため利用できることを提案する．我々の定義する情報量は，連続値属性Aにおいて，属性値A(e)が閾値θより大きいか小さいかを知ることにより，対応する事例がクラスC_1かC_2に属する可能性に関する情報を表す量であり，閾値情報量と呼ぶ．

On the Division of Continuous Values in the Construction of Decision Trees.

Hiroyuki IZUMI, Hideki SATO
FUJITSU LABORATORIES LDT.

以下，この閾値情報量を定義する．

訓練事例集合S_1,S_2における属性Aの値の頻度分布関数をそれぞれf(x),g(x) (x∈ [$minA_0$, $maxA_0$]) として

$$F(\theta)=\int_{minA_0}^{\theta} f(x)dx, \quad G(\theta)=\int_{minA_0}^{\theta} g(x)dx$$

とおくと，F(θ) は，Aの値が$minA_0$ からθまでの値をとるようなクラスC_1に属する事例の個数を与える．$F(maxA_0)=n_1$, $G(maxA_0)=n_2$ であり，分割表[2,3]は表1のようになる．

表1　閾値情報量のための分割表

	クラスC_1	クラスC_2	小計
$<\theta$	$F(\theta)$	$G(\theta)$	$F(\theta)+G(\theta)$
$\geqq\theta$	$n_1-F(\theta)$	$n_2-G(\theta)$	$n-(F(\theta)+G(\theta))$
小計	n_1	n_2	n 〔$=n_1+n_2$〕

これから，θで分割したときの相互情報量を，

$$\begin{aligned}I(A,C,\theta) &= (1/n)*\{F(\theta)*\log_2 F(\theta)+G(\theta)*\log_2 G(\theta)\\ &+(n_1-F(\theta))*\log_2(n_1-F(\theta))\\ &+(n_2-G(\theta))*\log_2(n_2-G(\theta))\\ &-(F(\theta)+G(\theta))*\log_2(F(\theta)+G(\theta))\\ &-(n-(F(\theta)+G(\theta)))\\ &\quad *\log_2(n-(F(\theta)+G(\theta)))\\ &-n_1*\log_2 n_1-n_2*\log_2 n_2+n*\log_2 n\}\end{aligned}$$

と定めることができる[2]．属性Aとクラス集合Cを固定して考えるとき，I(A,C,θ) をI(θ) と表し，属性Aとクラス集合Cに関する閾値情報量と呼ぶ．

閾値情報量I(θ_j) が極大値をとるような閾値群$\theta_1<\theta_2<\ldots<\theta_h$ を極大閾値群と呼ぶ．

区間作成において次のような閾値設定法を提案する．

「I(θ) を [$minA_0$, $maxA_0$] の範囲で計算して，極大閾値群$\theta_1<\theta_2<\ldots<\theta_h$ を算出し，[$minA_0$, $maxA_0$] をそれらで分割して有限個の排他的な区間 [$minA_0$, θ_1), [θ_1, θ_2),....,[θ_h ,$maxA_0$] を構成する．」

これ以後，極大閾値群を簡単に求め得る場合，および従来の簡単な閾値設定法との関係を考察する．

6．極大閾値の例

以下では，典型的な分布の場合の閾値情報量および極大閾値群について述べる．

6.1　分布が重ならない場合

図1のように，属性Aの値において，事例集合S_1とS_2との頻度分布が重ならないものとする．$x_1\leqq x_2$の仮定から$maxA_1<minA_2$ より，次の定理が成立する．

〔定理〕　$maxA_1<minA_2$ のとき，

$\theta_1\leqq maxA_1<\theta_2\leqq\theta_3<minA_2\leqq\theta_4$ となるような任意の$\theta_1,\theta_2,\theta_3,\theta_4$ に対し，

$I(\theta_1)<I(\theta_2)$, $I(\theta_4)<I(\theta_2)$, $I(\theta_2)=I(\theta_3)$

である．

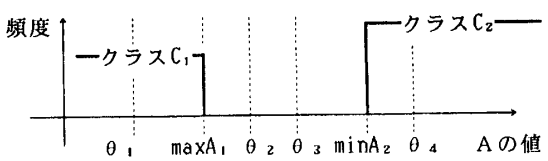

図1　分布が重ならない場合の閾値例

この定理は，$maxA_1<minA_2$ であるときには，クラスC_1とクラスC_2の訓練事例群がどのような頻度分布をしていても，$maxA_1<\theta<minA_2$ である任意の値θが極大閾値になることを意味する．

6.2　一様分布の場合

図2の実線のように，集合S_1,S_2 の属性Aの値域が重なりをもち，かつ一様分布に従っているとする．閾値情報量I(θ) は，図2の点線で図示したように，$minA_2$ と$maxA_1$ の部分で極大値をとる双峰型の曲線となる．

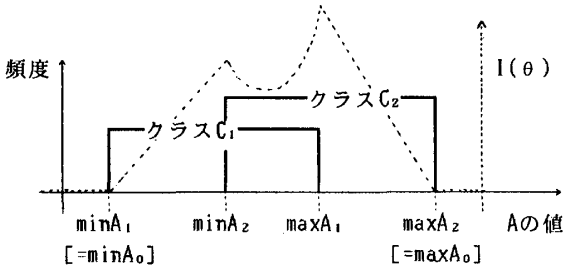

図2　一様分布と閾値情報量

それゆえ，訓練事例を一様分布とみなして，その極大閾値を閾値として採用することは，訓練事例の頻度分布の最大値や最小値をAの閾値とみなすことになる．

6.3　事例数がほぼ等しい正規分布の場合

下に示す定理は，S_1,S_2 が等分散の正規分布で同数個の要素を持つときに S_1,S_2の平均値x_1,x_2 の中間値$\theta_0=(x_1+x_2)/2$ でI(θ_0)が極値を取ることを示している．我々の実験によればI(θ_0)は最大値であることが確認された．

〔定理〕集合S_1とS_2の属性Aにおける頻度分布が，それぞれ正規分布$N(x_1,s_1)$と$N(x_2,s_2)$に従い，$n_1=n_2$ かつ $s_1^2=s_2^2$のときには，$I((x_1+x_2)/2)$は極値をとる．

S_1,S_2 が正規分布でほぼ同数個の要素を持つときに，それらの平均値の中間値を閾値として用いることは，閾値情報量をできるだけ大きくするように閾値を設定していることになる．

7．おわりに

決定木構成法において，連続値を有限個の区間に分割するための閾値設定の方法について議論した．そのために，閾値に関する情報量を定義し，その値が極大となるように区間分割を行うことを提案した．また，典型的な場合として，訓練事例の属性の頻度分布が重ならない場合，一様分布で重なりを持つ場合，訓練事例がほぼ同数で正規分布をなす場合について，相互情報量を極大にするような閾値がどこに存在するかについて述べた．

参考文献

〔1〕Quinlan,J.R.,"Induction of Decision Trees," Machine Learning, vol.1,pp.81-106(1986).

〔2〕Mingers,J.,"Expert Systems-Rule Induction with Statistical Data," J.Opl.Res.Soc.,Vol.38., No.1,pp.39-47(1987).

〔3〕本間鶴千代,"統計数学入門,"森北出版(1970).