決定木の構造検定

1X−06

高光 知哉　三浦 孝夫
法政大学工学部電気電子工学科

塩谷 勇
産能大学経営情報学部

概 要

決定木の評価法には誤分類率があるが、正確性以外で評価する方法は確立されていない。本論文では χ^2 検定を用いた決定木の構造検定について述べる。これを用いることで生成された枝の検定が可能になり影響力が高い枝を抽出することができる。また、検定を用いることで数学的な保障を与えることができる。

1　動機と解法

あるデータベースのデータを用いて決定木を作る。この決定木を信頼できると評価する方法を考える。考えられる方法として誤分類率が上げられる。この方法は枝別に正答率の評価をしている。だが、誤分類率以外で評価する手法は確立されていない。

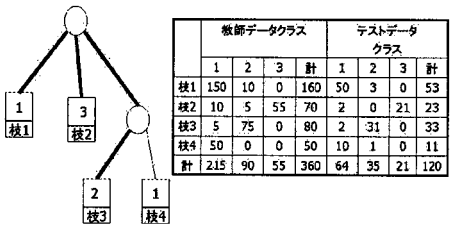
本論文では、決定木の構造検定を提案する。これは、正解数を見るだけでなく間違った分類まで考慮に入れて決定木の評価を行う。独立性の検定と累積 χ^2 検定を行うことで決定木全体の評価と順位付けを行う。各枝の順位付けというのは、決定木全体の中で影響力が強い順に選んでいくことである。影響力が強いということは、決定木の構成内で重要な役割を果たしていると考えられる。つまり、重要な枝の抽出していることになる。そして、各枝に対して適合度検定を行う。このように、検定という手段を用いることによって数学的に信用できる評価をすることができる。

2　構造検定

χ^2 検定を用いて決定木の構造検定を行う。ある教師データを用いて決定木を作成すると、4枝からなる決定木を作成できる。この木を用いてテストデータを分類する。このデータを用いて構造検定の説明を行う。図1にデータと決定木を示す。葉に書かれているのはクラスである。各枝には1〜4の番号を与える。表には各枝に分類された教師データとテストデータをクラス別にまとめたものを示す。

2.1　独立性の検定

独立性の検定は2つのカテゴリの関連性について調べる検定である。これを用いて枝とテストデータが分類されたクラスの関連性について調べる。i枝に分類されたデータのjクラスの数を y_{ij} として、その期待値 $E_{ij} = \frac{y_{i.}y_{.j}}{y_{..}}$ を計算する。$y_{i.}$ はi枝に分類されたデータの総数、$y_{.j}$ はjクラスを持つデータの総数、$y_{..}$ は総データ数である。この期待値を用いて独立性の χ^2 を $\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(y_{ij}-E_{ij})^2}{E_{ij}}$ で計算する。枝1のクラス1を例にとると、期待値は $E_{ij} = \frac{64\cdot 53}{120} = 28.26$ となり、χ^2 は $\frac{(50-28.26)^2}{28.26} = 16.71$ となる。これをすべての枝とクラス組み合わせについて計算を行い合計する。その結果、χ^2 は196.34となる。この値と χ^2 分布の値を比較する。自由度8、有意水準0.05の χ^2 分布の値は15.5となる。2つを比較すると計算した値の方が遥かに大きい。これより、枝と分類されたクラスの関係が深い。よって、この枝集合はデータを分類する機能を果たしている。もし χ^2 分布の値の方が大きかった場合、枝と分類されたクラスの関係はないことになる。よって、その枝集合は決定木としての機能を果たしていない。



	教師データクラス				テストデータクラス			
	1	2	3	計	1	2	3	計
枝1	150	10	0	160	50	3	0	53
枝2	10	5	55	70	2	0	21	23
枝3	5	75	0	80	2	31	0	33
枝4	50	0	0	50	10	1	0	11
計	215	90	55	360	64	35	21	120

図 1: データと作成した決定木の例

2.2　累積 χ^2 検定

累積 χ^2 検定は χ^2 値を分割することができる。また、どのように分割しても合計は変わらない性質を持つ。累積 χ^2 検定を用いて1つの枝とその他という形で分割する。このとき、分割した値が大きくなるような1つの枝を選択する。なぜなら、値が大きい枝が決定木全体の中で影響力が強いためである。分割したときの値を計算する式は $\chi^2(I_1;I_2) = y_{..}((\frac{1}{\sum_i^{I_1} y_{i.}}) + (\frac{1}{\sum_i^{I_2} y_{i.}}))^{-1}\sum_{j=1}^{m}\frac{1}{y_{.j}}((\frac{\sum_i^{I_1} y_{ij}}{\sum_i^{I_1} y_{i.}}) - (\frac{\sum_i^{I_2} y_{ij}}{\sum_i^{I_2} y_{i.}}))^2$ である。例えば、独立性の検定で出した値を枝別に分類する方法として「2と1と3,4」「1,2と3,4」等がある。「2と1とそれ以外」の分割を例にすると、最初に2とそれ以外で分割を行う。計算は $\chi^2(2;1,3,4) = 120(\frac{1}{23} + \frac{1}{(53+33+11)})^{-1}\frac{1}{64}(\frac{2}{23} - \frac{50+2+10}{53+33+11})^2\frac{1}{35}(\frac{0}{23} - \frac{3+31+1}{53+33+11})^2\frac{1}{21}(\frac{21}{23} - \frac{0+0+0}{53+33+11})^2 = 107.49$ となる。その後、同様に「1と3,4」「3と4」の分割の計算を行う。分割した値の和は分割する前の χ^2 値に等しい。値が大きくなるように枝を分割した結果「2と3と1,4」となる。それぞれの値は107.49、88.79、0.06である。これらの値の和は196.34となる。これは分割する前の χ^2 値である。よって、影響力の高い枝は2、3、1と4の順になる。

2.3　適合値検定

適合度検定では各枝でのクラスごとの分布を調べ、誤りを含む教師データとの対応付けを評価する。各枝の教師データ

Testing Structure of Decision Trees
Tomoya Takamitsu, Takao Miura, Isamu Shioya
Hosei University, Dept.of Elec. and Elec. Eng.
Kajino-cho 3-7-2, Koganei, Tokyo, JAPAN
Sanno University, Dept.of Management. and Info.
Kamikasuya 1563, Isehara, Kanagawa, JAPAN

自由度は (枝数 − 1) × (クラス数 − 1) で定義される。

のクラス分布を理論値として $P_{ij} = \frac{y_{ij}}{y_{i.}}$ で計算を行う。そして、テストデータを式 $\sum_{j=1}^{m} \frac{(y_{ij}-y_{i.}P_{ij})^2}{y_{i.}P_{ij}}$ に代入して $\mathcal{X}^2$ 値を求める。その結果分布が一致しているならば、その枝は誤分類を考慮に入れて正常に動いている。枝 1 を例にとると $P_1 = \frac{150}{160} = 0.937$、$P_2 = \frac{10}{160} = 0.062$、$P_3 = \frac{0}{160} = 0$ となる。P_3 が 0 なので、0.001 に補正する。$\mathcal{X}^2$ 値は $\frac{(50-53\cdot 0.937)^2}{53\cdot 0.937} + \frac{(3-53\cdot 0.062)^2}{53\cdot 0.062} + \frac{(0-53\cdot 0.001)^2}{53\cdot 0.001} = 0.084$ となる。同様に枝 2、3、4 の $\mathcal{X}^2$ 値も計算するとそれぞれ、2.62、0.035、89.022 となる。この値と $\mathcal{X}^2$ 分布を比較する。$\mathcal{X}^2$ 分布の自由度は 2、有意水準は 0.05 とすると値は 5.99 である。よって、分布を下回っている枝 1、2、3 は分布が一致しているので信用できる。分布の一致していない枝 4 は信用できない。たとえ誤分類率が 0 に近くても、枝が意図している分布通りに動いてないので信用できない。

2.4 検定の流れ

1. 独立性の検定 (決定木の内部構造の検定)
2. 累積 $\mathcal{X}^2$ 検定 (影響力の高い枝の抽出)
3. 適合度検定 (各枝の信頼性の評価)

以上の手順で決定木の構造検定を行うことができる。

3 実験

3.1 実験データと手順

本実験には Car Evalution Database[3] を用いる。このデータは 1728 台の車のクラス属性（unacc,acc,good,v-good）と 6 種類の車に対する評価を属性として持っている。この 1728 件のデータを 8:2 の割合で分割して教師データ:テストデータとし、決定木を作成する。本実験では 1383 件を教師データ、345 件をテストデータという形で分割を行う。決定木作成は、C4.5 と同様にエントロピに基づいて行う。これを用いて決定木を作成すると、57 枝からなる決定木を作成する。作成した枝には 0 〜 56 の枝番号を与える。作成した決定木を用いてテストデータを分類し、その結果を用いて決定木の構造検定を行う。

3.2 実験結果

テストデータが入った 41 の枝を用いて、独立性の検定を行うと、$\mathcal{X}^2$=632.677 となる。この値を $\mathcal{X}^2$ 分布と比較を行う。比較する $\mathcal{X}^2$ 分布の自由度は 120、有意水準は 0.05 とする。$\mathcal{X}^2$ 分布の値は 146.567 である。$\mathcal{X}^2$ 値 $>$ $\mathcal{X}^2$ 分布値となっているので、この枝集合は機能を果たしている。独立性の計算結果を用いて、累積 $\mathcal{X}^2$ 検定と適合度検定を行う。累積 $\mathcal{X}^2$ 検定の結果は図 2、図 3 のルール番号の部分に示す。ここの番号の順番が抽出したルールの順番である。また、図 3 に分割した枝の $\mathcal{X}^2$ 値合計が全体の $\mathcal{X}^2$ 値の何%を示しているかをグラフで示す。図 2 に各枝に分類されたテストデータをクラス別にまとめたグラフを示す。適合度検定の結果は図 3 に示す。各枝の適合度検定の計算した $\mathcal{X}^2$ 値と比較する $\mathcal{X}^2$ 分布の値を示す。比較する $\mathcal{X}^2$ 分布の自由度は 3、有意水準は 0.05 とする。また、比較するために各枝の誤分類率を図 3 に示す。また、各図のグラフに線が入っている部分がある。これは、線より右側の枝はテストデータは枝に分類されたが累積 $\mathcal{X}^2$ 検定で選ばれなかった枝を表している。

4 評価

4.1 実験評価

図 3 を見ると、最初の 10 の枝で全体の $\mathcal{X}^2$ 値の 70%を占めている。よって、最初の方に選ばれた枝は非常に影響力が高いことがわかる。適合度検定結果を見てみると、大体の枝は $\mathcal{X}^2$ 分布の値より下回っているため正常に動いていると確認できたが、いくつかの枝は信用できないと判断できる。グラフより、影響力が高い枝でも信用できない枝が存在することがわかる。この信頼できない枝は誤分類率から見ても信頼

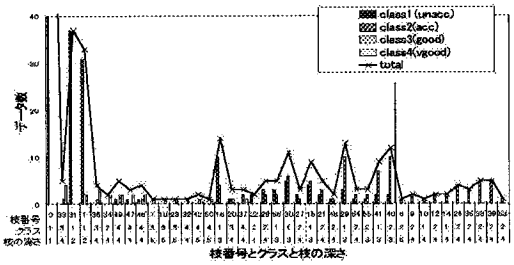


図 2: 各クラスに分類されたデータ数

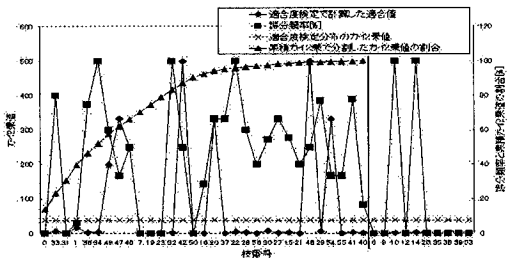


図 3: 誤分類率と適合度検定結果の比較と $\mathcal{X}^2$ 値の割合

に置けない、つまり見方を変えても信頼できないことがわかる。適合度検定で正常に動いていると確認がとれても、誤分類率は低いわけではない。これは、判断する基準が違うためである。図 2 をみてみると、選ばれる枝に偏りが存在していることがわかる。1 つ目は浅い枝が選ばれやすい。よって、極端に浅い枝は影響力が強いと判断できる。2 つ目は特定のクラスのデータが入っている枝が選ばれやすい。データを見てみると、good,vgood が入っているクラスは影響力が高いものが多い。逆に acc は選ばれにくい傾向がある。よって、acc のクラスのデータは決定木作成時にあまり重要な役割を果たしていないことがわかる。

4.2 構造検定評価

この評価法はテストデータが分類されなかった枝については評価ができない。よって、すべての枝を評価するには、すべての枝にデータが行き渡るようなテストデータを用意しなければならない。そういう意味で、テストデータに対する依存性が高い評価法である。また、枝の数が多いとすべての枝を抽出する前に累積 $\mathcal{X}^2$ 検定が終わってしまうことがわかった。しかし、後半になればなるほど影響力が小さい枝ばかりなので打ち切ってしまっても問題はない。

5 結び

本実験の結果では枝の選びかたに偏りが見られることがわかった。これは、属性やクラスの偏りが関与している可能性がある。また、教師データの正当性についても考える必要がある。おかしなデータであればしっかりとした決定木を作ることができない。これらを意識した決定木の作成や評価について考える必要がある。

参考文献

[1] 東京大学教養学部統計学教室: 自然科学の統計学, 財団法人東京大学出版会 (1992)

[2] 古川康一: AI によるデータ解析, 株式会社トッパン (1995)

[3] Marko Bohanec: Car Evalution Database, UCI(http://www1.ics.uci.edu/ mlearn/) (1997)