偏りのあるデータに適用可能な決定木学習アルゴリズムの検討

小暮　陽太†　　木村　昌臣†

芝浦工業大学†

1. 初めに

決定木学習は過去の事例からその事例に対する一般的な規則を導き出す帰納学習と呼ばれる手法一つであり、その分類ルールが明確であることから分類や予測を行うデータマイニング手法として広く用いられている。しかしこの手法には，学習データに少数の事例のみが属するクラスがノイズとして扱われてしまい，そのようなクラスは重要であっても無視されてしまう問題点がある．そこで本研究では属する事例数が少ないクラスが存在する場合にでも適用可能な決定木学習アルゴリズムの提案を行う．

2. 手法

サンプリング法[1]はクラスに属するレコード数が均等になるようにデータからサンプリングを行い，そのサンプルを学習データとして用いることで決定木学習を行う．そして得られた決定木の誤分類率が高い部分を修正するように再サンプリングを実施する手法である．しかし，この手法は事例数の多いクラスの情報が欠落し解析結果に反映されないという問題が生じる．

そこで本研究では最もレコード数が少ないクラスに合わせてデータを分割し，それらを組み合わせて得られる偏りのないデータ全てから全体の情報を反映した分類ルールを導出する手法を提案する．これにより事例数が少数のクラスの情報をノイズとして扱わずに，かつ事例数が多いクラスの情報を削らずに決定木学習を行うことができる．

2.1. サンプリング

既存研究であるサンプリング法ではレコード数を均等にするために，最もレコード数が少ないクラスに合わせてサンプリングを行った．しかしその手法では抽出されないレコードがあるため，学習データ全体を反映しない結果が出力されてしまう可能性がある．

そこで本研究では最もレコード数が少ないクラスに合わせて決定木の各ノードに所属するデータを分割する．これによりノードに属する各クラスは，最もレコード数が少ないクラスのレコード数を1単位とした複数のグループに分けることができる．

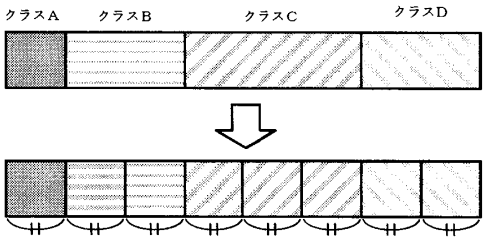

図1　最小のクラスに合わせた分割

各クラスからこのグループを1つずつ抽出することで得られる各組み合わせをサンプルとして用いる．

2.2. 実験計画法に基づく組み合わせの削減

最もレコード数が少ないクラスと他のクラスのレコード数に大きな差がある場合，組み合わせの数が膨大になってしまう．

そこでAll-pair法を用いることで組み合わせ数を削減にする工夫を行った．All-pair 法は実験計画法の一手法で2変数間の網羅を行うことで実験回数を削減する手法である．

2.3. 提案手法による決定木の成長

各サンプルに対して情報量利得比を用いて分割を行う説明変数を導出する．情報量利得比とは決定木学習アルゴリズム C4.5[2]で用いられる分割の良さを表す指標で，情報の煩雑さを示すエントロピーに基づいて計算される．こうして導出した分類ルールはサンプル特有のレコードの影響でノード全体のデータの情報を反映していない可能性がある．そこでサンプルを分類ルールで分割した際の最大の事例を持つクラスをそれぞれの分割の正解とし，他の組み合わせのサンプルを用いて正答率を計算する．

以下が提案手法における決定木の成長の具体的な流れになる．N(t)をノードt内の事例数，P(c|t)をノードt内のレコードがクラスcに属する割合とし，学習データはn個のクラス$c_1,\cdots,c_n$を持ち，m個の説明変数$a_1,\cdots,a_m$を持つとすると提案アルゴリズムの実行手順は以下の通りである．

① 学習データ全体をルートノードとし，Tとおく

② ノードTの内，属するレコード数が最小のクラスをC_{min}とする

Decision tree Algorithm to be applied to unbalance-datasets.

†Yota Kogure, †Masaomi Kimura

†Shibaura institute of technology

③ 全てのクラス$c_1,\cdots,c_n$のレコードを

$$L=\frac{N(C_i|T)}{N(C_{min}|T)} \quad \cdots(1)$$

個のグループに分割し，クラスc_1のグループをそれぞれ$g_{i1},g_{i2},g_{i3},\cdots,g_{iL}$とする

④ 全てのクラスから一つずつグループを取り出す組み合わせをそれぞれ$G_1,G_2,G_3,\cdots,G_r$とする

⑤ G_s 説明変数a_jを用いて分割し，ノードTの子ノードとする

⑥ k個の子ノード$t_1,\cdots,t_k$についての情報量利得比を以下の式で計算する

$$\frac{\log(n)-\sum_{l=1}^{k}\frac{N(t_l)}{N(G_s)}\left(-\sum_{q=1}^{n}P(c_q|t_l)\log P(c_q|t_l)\right)}{-\sum_{l=1}^{k}\frac{N(t_l)}{N(G_s)}\log\left(\frac{N(t_l)}{N(G_s)}\right)} \quad \cdots(2)$$

⑦ 情報量利得比が最大となる説明変数a_{min}をG_sの選択とし，G_sをa_{min}で分割しノード T の子ノードとする．

⑧ k個の子ノード$t_1,\cdots,t_k$について最大のレコードを持つクラスを各ノードの正解とし，正答率を全ての組み合わせ$G_1,G_2,G_3,\cdots,G_s$について求める

⑨ ⑤～⑧を全ての説明変数について行う

⑩ ④～⑨を全ての組み合わせに対して行い，全ての説明変数についての正答率の合計値が最も高い変数をノードTの分類ルールとして適用する

⑪ ノードTの子ノードのうちの一つをノードTとする

⑫ 全ての子ノードに対しこれ以上分割できなくなるまで②～⑪を再帰的に行う

2.4. 剪定

多くの決定木学習アルゴリズムと同様に本研究の提案手法では，ノードを決められた指標がこれ以上よくならなくなるまで分岐を繰り返す．こうしてできた決定木は学習データ特有の特徴までを反映してしまうという，過学習が行われている可能性が高い．そこで本研究でも学習用データ以外の評価用のデータを決定木に適用し，その誤分類率を調べることで剪定を行う．

剪定を行うかどうかの判定はそのノードの分類を剪定した場合としなかった場合の決定木全体の誤分類率を比較し，剪定によって誤分類率が減少する場合に剪定を行う．

これを葉ノードから順に上位のノードに対して行うことで決定木を剪定する．

3. 評価実験

提案手法を実装したプロトタイプシステムを用いて，提案手法の評価を行った．

決定木学習を行う対象データとして独立行政法人医薬品医療機器総合機構で収集された医薬品投与ヒヤリハット事例第1回～第14回の，全2259件をもとに作成されたデータを使用した．うち「主要因」を目的変数に「事例内容」「チェック通過の有無」「名称類似度」「剤型違い」「規格量違い」を説明変数とした．また全データのうち 60%を学習データ,20%を評価データ,20%をテストセットとして用いた．また実験は提案手法と C4.5 アルゴリズムに対して5回試行し，その結果を比較した．

表1 提案手法と C4.5 での誤分類率

実験 No.	提案手法	C4.5
1	48.48%	49.57%
2	45.89%	46.32%
3	47.84%	47.62%
4	46.32%	47.19%
5	44.81%	47.84%

提案手法では C4. 5アルゴリズムと比べ決定木全体の誤分類率が 5 回中 4 回上昇した．またその誤分類率の差の平均値は 1.04%であった．

C4.5 アルゴリズムではルートノードの分類は「事例内容」であったのに対して提案手法では「チェック通過の有無」 が選出された．これは提案手法でレコード数の少ないクラスが重要視されたことにより，それらのクラスを分類する説明変数がルートノードの近くで導出されるようになったためであると考えられる．これらの場合の各クラスのレコードについてのエントロピーの総和を求めたうえランダムに選ばれた場合の値で標準化したところ「チェック通過の有無」が「事例内容」に比べて低い値となった．これは本手法の結果と両立する．

4. まとめと今後の課題

提案手法によってつくられた決定木の全体の誤分類率が既存手法に比べ高いものであり，少数のクラスを導出しやすいものであることがわかった．

今後，本研究の提案手法を他のデータセットに適用して，その適用範囲や有効性について調べる必要がある．

参考文献

[1] Chawla, V. N. et al. : SMOTE Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, vol. 16, pp. 321-357 (2002).

[2] Quinlan, J. R. : C4.5: Programs for Ma-chine Learning. Morgan Kaufmann (1993).