重み付き分類規則による保健データからのデータマイニング

高江 徹[†]　近棟 稔[†1]　有村 博紀[†]　篠原 歩[†]　井上 仁[‡]　武谷 俊一[‡2]
1Y−8　上園 慶子[§]　川崎晃一[§]
[†]九州大学大学院システム情報科学研究科 [‡]九州大学中央計数施設 [§]九州大学健康科学センター

1. はじめに

分類規則の発見は，データマイニングにおける主要な問題の1つである．とくに，未知データの予測だけでなく，データの特徴をとらえる上でも有効である．分類規則の表現方法の1つとして，各ノードに分類規則をもつ決定木は実用的な観点から広く用いられている．Quinlan による C4.5 システム[5]は代表的な決定木生成アルゴリズムであるが，相関の強い数値データに関しては，非効率的である．これに対して，福田等[1]は，数値データの組からなる平面の部分領域を分類規則として用いて，小さい決定木を高速に求めるアルゴリズムを提案している．

このアルゴリズム (DT) は，各ノードで，全ての数値属性の対ごとに，最適2次元分類規則を計算し，その中から最も値のよい分類規則を採用し，残りを全て破棄する．各ノードごとに多数の分類規則の候補生成を繰り返すために，分類規則の候補が大量に存在する場合，決定木構築にかかる計算時間は膨大になる．

本稿では，決定木の根ノードを構築するのと同等の計算時間によって生成可能な「重み付き分類規則」を提案する．重み付き分類規則 (Weighted Aggregation Classifiers, WA) とは，決定木の根ノード生成時に候補に挙がる全ての分類規則に，エントロピーを基準とした重みを付け，全ての属性の重み付き多数決によって予測を行う分類規則である．このため，WA 規則は属性全体を考慮することで，明確な法則が出にくい対象や，例外値やノイズを含む不完全データに対しても，うまく働くと期待できる．

そこで，本稿では，重み付き分類規則を生成するアルゴリズムを実装し，決定木生成アルゴリズム[1]と実験による比較を行う．

2. 重み付き分類規則

重み付き分類規則 H は，4つ組の集合 $H = \{(r_i, c_i^1, c_i^0, w_i) \mid i = 1, \cdots, k\}$ である．各 $i = 1, \cdots, k$ に対して，r_i は分類規則であり、w_i は r_i の分類精度を表す重みである．各 c_i^1 (c_i^0) は，入力タプル $\mathbf{t}$ が r_i を満たす (満たさない) ときに，目標属性が真である確信度 (条件付確率) である．

入力タプル t に対して、H による目標属性の予測値は，すべての分類規則の重み付き多数決で決める:

$$H(\mathbf{t}) = \left[\sum_i w_i \cdot conf_i(\mathbf{t}) \geq \sum_i w_i \cdot (1 - conf_i(\mathbf{t}))\right].$$

ここに $\mathbf{t}$ が r_i を満たすときは $conf_i(\mathbf{t}) = c_i^1$，それ以外のとき $conf_i(\mathbf{t}) = c_i^0$ と定義する．$[P]$ は、命題 P が真 (偽) ならば 1(0) を返す関数である．

分類規則の候補として，離散属性に関する規則 "$t[A] = v$"[5] に加えて，福田等[1]で提案された1次元区間 I を用いた規則 "$t[A_1] \in I$" と，2次元直交凸領域 R を用いた規則 "$(t[A_1], t[A_2]) \in R$" を採用する．離散属性に関してはすべての値 v について，数値属性に関しては，各属性 (の組み合わせ) すべてに関して，福田等[1]の最適化手法でエントロピー値 $Ent(r_i)$ (後述) を最小化する規則 r_i を一つづつ選びだし，重み付き分類規則に用いる．

分類規則 r_i は，その予測値によってタプル集合 S を2つの部分集合 S_1, S_0 に分割する．S_i の例全体の数を N_i とおき，これに対する正例の比率を $0 \leq p_i \leq 1$ とおく $(i = 0, 1)$．分割の良さは関数 $\psi(p) = -p\log p - (1-p)\log(1-p)$ ではかり，エントロピー関数を $Ent(r_i) = \psi(p_1)N_1 + \psi(p_0)N_0$ とする．最後に，分類精度の良い規則に大きな重みを付けるために，重みを $w_i = \max_i\{Ent(r_i)\} - Ent(r_i)$ で定義する．

3. 実験

本節では，WA アルゴリズムと DT アルゴリズム[1]を実装し，予測精度と学習時間を比較する実験を行った．実験データには，標準的データとして UCI repository のデータ[4]と，実データとして大規模な定期健康診断データ (以下，定健データと略す)[2]を用いた．実験は，Unix ワークステーション (Ultra Sparc IIi 300MHz, Solaris 2.6) 上で実施した．

Knowledge Discovery from Health Data Using Weighted Aggregation Classifiers
Toru Takae, Department of Informatics, Kyushu University, Hakozaki 6-10-1, Fukuoka 812-8581, Japan.
[1] 現在，三菱電機 (株)．[2] 現在，九州大学アドミッションセンター．

表 1 UCI レポジトリデータセット [4] における WA と DT の分類精度 (%) と計算時間 (sec) の比較．分類精度が良い方を下線で示した．

Dataset	size	Acc_{base}	WA Acc	WA Time	DT Acc	DT Time
Breast Canser	699	65.52	<u>97.51</u>	237	95.61	666
Liver Disorder	345	57.97	<u>67.83</u>	51	50.46	202
Pima Diabetes	769	65.10	69.40	234	<u>69.92</u>	1216
Balance Scale	625	53.92	<u>85.59</u>	31	79.36	106
Titanic	2201	67.70	77.60	0.5	<u>79.05</u>	2

表 2 国立大学学部生の定健データ [2] における WA と DT の平均および最大分類精度 (%) の比較

属性	Acc_{base}	WA Acc_{avr}	WA Acc_{max}	DT Acc_{avr}	DT Acc_{max}
実験 1 (BMI)	60.00	59.51	63.42	60.11	63.93
実験 2 (BMI+Others)	60.00	70.71	74.13	62.92	70.09

3.1 UCI レポジトリデータ

最初に，UCI repository[4] の 5 つのベンチマーク用データセットで実験を行った．実験方法としては，まずデータセットをランダムに 2 つの集合に分割し，一方を学習用の訓練例とし，もう一方を仮説の分類精度を評価するためのテストデータとして実験する．その後，訓練例とテストデータを交換し，同じ実験を行う (2-fold cross varidation)．実験値としては，この 2 度の実験の平均を取った．

表 1 にその実験結果を示す．Acc_{base} は，自明な分類精度，すなわち，正例または負例の割合の大きい方を示す．実験結果から，重み付き分類規則は分類精度がわずかに高く，学習時間が短い事がわかる．

3.2 定健データ

次に，実データセットである大規模定健データを用いて，両方のアルゴリズムの比較を行った．

データセット． データセットは，国立大学等保健管理施設協議会による定健データ [2] である．このデータセットは，1995 年に全国の国立大学 95 校で実施された健康状態に関する大規模な実態調査で得られたものである．これは，100 以上の離散および数値属性に関する 30 万超のレコードからなる定健データである．

実験方法． 実験方法としては，まず収縮期血圧 (SBP) を目標属性とし，肥満度指数 (BMI) を示す数値属性と生活習慣に関する 13 個の順序付離散属性を説明属性とした．原データから 75 個の訓練例と 3653 個のテストデータをランダムサンプリングし，実験を行った．100 回の試行を行い，分類精度の平均値と最大値を計算した．なお，データの 60%は正例なので，その自明な分類精度は $Acc_{base} = 60\%$ である．

実験結果． 最初に，典型的な解析として，BMI のみを説明属性として用いて高血圧 ($SBP \geq 140$) の予測を試みた (実験 1)．次に，多数の属性を用いた解析として，BMI と全ての生活習慣に関する属性を用いて高血圧の予測を試みた (実験 2)．表 2 に結果を示す．

実験 1 と実験 2 の結果を合わせると，アルゴリズム WA, DT の両方で，BMI と全ての生活習慣属性を用いた予測の方が，BMI のみを用いた予測よりも分類精度が良いことがわかる．BMI のみの方の分類精度は自明な精度 60%をわずかに上回る程度である．

実験 2 では，分類精度の平均値と最大値の両方で，WA が DT より予測精度が良いことがわかる．原データでは BMI と SBP は無相関に近く予測が難しいデータだが，WA は，自明な精度 60%に対して，74%の高い精度をもつ規則を見つけている．

4. まとめ

本稿では，重み付き分類規則を提案した．比較実験から，重み付き分類規則は決定木に比べて学習時間が短く，予測精度も高いという結果を得た．

参考文献

1) T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Constructing efficient decision trees by using optimized numeric association rules, In Proc. the 22th VLDB Conference, 146–155, 1996.

2) 国立大学等保健管理施設協議会, 学生の健康白書作成に関する特別委員会編. 学生の健康白書 1995, 基本編と応用研究編, 1997.

3) R. C. Holte. Very simple classification rules perform well on most commonly used datasets, Machine Learning, 11, 63–91, 1993.

4) P. M. Murphy and D. W. Aha. UCI repository of machine learning databases, 1994. `http://www.ics.uci.edu/mlearn/MLRepository.html`.

5) J. R. Quinlan. C4.5: Programs for machine learning. Morgan Kaufmann, 1993.