マルチメディア通信と分散処理 111－14
コンピュータセキュリティ 20－14
（２００３．２．２７）

Web識別問題における2クラス，多クラス決定木の比較

米山 計昇 * 高崎 勲 * 菊池浩明 * 中西 祥八郎 *

概要: Web 情報検索において Web ディレクトリ構造型の有効性が認識されている．しかし, 正確に Web ページをディレクトリ型に分類するには, 人間の主観に頼った選別が主流である．そこで, 我々は Web ページから抽出されたキーワードの集合に対し, 決定木アルゴリズム ID3 を適用することを試みる．c 個のクラスへ分類するには,2 クラス木と多クラス木の 2 通りのアプローチがある．多クラス木は, c 種の葉を許した木で, 単一の木で識別を実行する．2 クラス木は, 葉が「+」「-」のどちらかであり, c 個の木で一つの識別を行う．本稿では, この 2 クラス, 多クラス決定木の性能や精度について報告する．

Comparison of Binary and Multiple-classes Decision Trees for classification of Webpage

Kazunori Yoneyama * Isao Takasaki * Hiroaki Kikuchi *
Shohachiro Nakanishi *

Abstract: A web directry is useful service for retrieving webpages. However a classification of webpages is commonly made by many human opperator's subjective. In order for automatic classification of webpages, given a set of keywords extracted from webpages we apply to a decision tree learnig algorithm including the ID3 algorithm. There are two possible approaches to classify webpages into c classes - a binary tree and a multiple-class tree. A multiple-class tree is of c diferent kind leaves and classifies webpages by just one tree. A binary tree is a tree in which leaves are '+' or '-'. A set of c independent trees gives an integretedd classification of webpages. In this paper we estimate these performance and accuracy for some training data.

1 はじめに

Web 情報検索において,Web ディレクトリ構造型の有効性が認識されている．しかし, 正確に Web ページをディレクトリ型に分類するためには, まだまだ人間の主観に頼った選別が主流である．この人的コストが, 新規ページの Web ディレクトリへの更新を困難にしていた.

そこで, 我々の研究では,Web ページの自動分類を試みる．機械的に分類を行う手法に,ID3 や C4.5 やニューラルネットワークがあるが, 論理決定木生成アルゴリズムの一つである ID3 を利用している．

Web ページの識別は, 次の様に実行する．学習データ (事例) は, いくつかの Web ページとする．Web ページには, 非常に多くのキーワードが存在する．そこで, 識別に有効なキーワードを絞り込む評価測度として, TF-IDF [5] を使用し，キーワード数を絞り込んで, これを学習データとする．その学習データに ID3 を適用させ,2 クラスと多クラスの決定木を生成した．そしてテストデータをそれぞれの決定木に適用させる事で, 決定木のサイ

*東海大学大学院工学研究科，〒 259-1292 神奈川県平塚市北金目 1117，Graduate school of Engineering, Tokai university, 1117 Kitakaname, Hiratsuka, Kanagawa, 259-1292, Japan {kazu, issa, kikn}@ep.u-tokai.ac.jp

ズと分類誤差率を求める.

この識別問題に対し，我々は抽出キーワード「沖縄」についての Web ページの自動分類を試みて, 実験結果を報告した [1]. しかし，作られた決定木に, 大きな誤差を含む結果となった. その要因が, 学習データそのものに, 人間の主観による選別の曖昧さがあったためである. そこで, 本稿では, yahoo[6] のディレクトリ構造を学習データとして採用した.

本稿では,2 クラス決定木と多クラス決定木のどちらが良い識別を与えるかを考える. 「良い識別木」とは, 分類誤差と決定木のサイズの両方を小さくする木の事である. 誤差は, 本来分類されるはずのクラスに分類されず, 間違ったクラスに分類されてしまった事例の総数である. サイズは, 生成された決定木の根や葉を含む全ての節点の数である. 我々は決定木生成の際に,2 値クラスと多値クラスの決定木を生成し, どちらが分類に優れているかを比較する. 2 クラス決定木は, 一つのクラスに対して一つの木を生成していくため，多クラス決定木よりも優秀であると予測される.

2 準備

2.1 Web ページ識別問題

本稿では, p 個の Web ページを事例の集合とし, それらに出現する k 個のキーワードを属性として扱う. あるページは, k 次元のブールベクトル $x = (0, 1 \ldots, 0) \in \{0,1\}^k$ によって一意に c 通りのクラス $C = \{1, 2, \ldots, c\}$ へ識別される. 正しい識別を与える写像を, A で表す. すなわち, $A : \{0,1\}^k \rightarrow C$ と表せる. Web ページ識別問題は, Web ページの部分集合が与えられた時に, A に対して誤差を最小化するアルゴリズム M を求める問題である.

2.2 誤差とサイズ

a) 誤差:

誤差には 2 通りある. 表 1 の例を考えよう. a は $+$ に属している事例が, 正しく $+$ に分類された数を示す. b は, $+$ クラスの事例が $-$ として識別されてしまった数, c は, $-$ クラスの事例が $+$ として識別されてしまった数の各々を表す. 本稿では, この 2 つの誤差の総和を誤差と定義する.

アルゴリズム M による x で特徴付けられるページの識別誤差を $\mathrm{err}(A(x), M(x))$ で表す. $\mathrm{err}(a, b)$ は $a = b$ の時 1, $a = b$ の時 0 を返す.

これを用いて, M の誤差は,

$$E(M) = \sum_{x}^{p} err(A(x), M(x)) = b + c \qquad (1)$$

と表記出来る.

表 1: 誤差表 サンプル

A\M	+	−
+	a	b
−	c	d

b) サイズ:

決定木は根 (ルート) から始まり, 節点 (ノード) を経由して葉 (リーフ) に至る. リーフには識別されるクラスが記述される. 木 T サイズ $S(T)$ は, 根・葉を含む全ての節点の数と定める.

2.3 2 値クラスと多値クラス

A,B,C の 3 つのクラスへ分類する問題を考えよう. 図 1 の木 T_M は, この識別を与える一つの例である.

一方, この問題を各クラス毎へと分解し, そのクラスへ属すか否かという 2 値の識別を与える 3 つの木 T_A,T_B,T_C によって解く方法も考えられる. 前者を多値木 (Multiple-class tree), 後者を 2 値木 (Binary tree) と呼ぶ. T_A と T_M の様に, 一般に 2 値木の方が多値木よりもサイズが小さくなる傾向があり, 最適化もかかりやすい. その反面, 多値木の c 倍の数の木が必要であり, 複数の木の出力が矛盾する可能性も生じる.

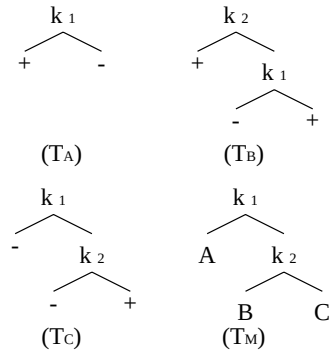

図 1: 2 値木 (T_A,T_B,T_C) と多値木 T_M の例

2.4 属性減少のための TF-IDF [5]

TF(Term Frequency) はある文章中に出現するキーワードの頻度である. この値が大きくなるほど, このキーワードが頻繁に出現することを表す. しかし,TF はキーワードの網羅性を高めるためには貢献するが, キーワードの特定性については必ずしも役に立たない. そこで, あるキーワードがどの程度その文章に特徴的に現れているかという特定性を考慮するために IDF(Inverse Document Frequency) がある. IDF は, あるキーワードが全文書中でどれくらいの文章に出現するかを表す尺度である. i 番目のキーワードに対する tfidf 値を

$$tfidf_i = \frac{1}{M}\log_2(\frac{M}{H_i})\sum_j^p \frac{\log_2(t_{ij}+1)}{\log_2 w_j} \quad (2)$$

と定める. ここで, p を事例の総数, H_i を i 番目のキーワードのヒット数. w_j を j 番目の事例を構成するキーワードの数, t_{ij} を j 番目の事例における i 番目のキーワードの数とする. 各々の単語に対して, tfidf 値で全てのキーワードに順位付けを行う.

2.5 信頼度と重み

ID3 や C4.5 では, 次に定められている信頼度と最小クラス数に基づいて, 不用な枝を取り除く簡単化 (枝刈り) を行う.

a) 信頼度 cf:

信頼度 cf(certainty factor) は決定木の枝刈りにを定めるパラメータである. 1%〜100%の値を取り, 小さな値ほど多くのノードを削除し, 木を簡単化するが, 誤差を大きくする.

b) 最小事例誤差 min:

最小事例誤差とは, 事例がクラスに分類される際に再帰を許すクラス数を定めている. 例えば,min を 2 に設定すると, 一つの葉に 2 つ以上のクラスの違う事例を分類しないという事を表している.

テストデータについて, これらを変動させその結果を図 2 に示す.

3 評価実験

3.1 実験目的

2 値決定木と多値決定木を生成し, どちらが, 良い決定木を与えるかを明らかにする.

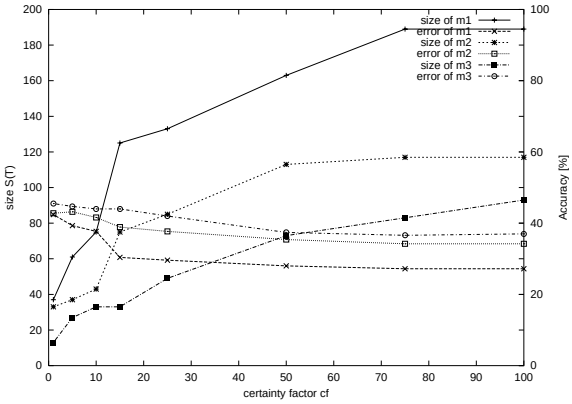

図 2: 枝刈りによるサイズと誤差の大きさ

3.2 実験方法

1. ディレクトリサービスより, あるカテゴリーとそのサブカテゴリークラスに属すページを全て取得する.

2. 1 のページから, 形態素解析ツール chasen により, 「名詞」のみを抽出し, それを属性とする. 意味のない語句を排除させるために, 数・接尾・接頭・感動・自立・代名詞・不変化・助詞となる名詞は除外する.

3. TF-IDF を利用し, 属性の数を k=250 まで絞り込む. つまり, $p = 515$,$k = 250$ の学習データと評価用のテストデータを作成する.

4. 3 のデータから, ランダムに 2 度サンプリングして, p 個のページの学習データと評価用のテストデータを作る.

5. 4 の学習データを ID3 にかけて, 多値クラス T_M とサイズ $S(T_M)$ を求める. T_M を用いて 4 のテストデータを識別し, 誤差 $E(T_M)$ を求める.

6. 5 の学習データを各クラス毎に 2 値化して, c 個の 2 値決定木 $T_1, \ldots, T_c$ を求める. (例えば, クラス B の 2 値化は,B を'+', B 以外を全て'-' へと割り当てる). 各木のサイズ $S(T_1), \ldots, S(T_c)$ と,5 と同じテストデータについての誤差 $E(T_1), \ldots, E(T_c)$ を求める.

7. 結果の信頼性を上げるため, 4 のサンプリングを変えて, 5,6 を数回繰り返す.

表 2: クラスと大きさ (2002/12 調べ)

クラス	カテゴリ	ページ総数	サンプル数
A	アマチュア	178	89
B	イベント	116	58
C	コラム, その他	23	12
D	障害者サッカー	19	10
E	選手	68	34
F	ナショナルチーム	39	20
G	フットサル	76	38
H	メディア	23	12
I	リーグ	484	242
	合計	1026	515

3.3 実験結果

yahoo[6] より, 「ホーム > 趣味とスポーツ > スポーツ > サッカー」というカテゴリを使用した. 使用サブカテゴリは, そのカテゴリが含んでいるページ数について表 2 に上位 9 カテゴリを対象とした. 表に示される 515 個のページから, 実験方法 4 に従って, $p = 257$ の学習 (トレーニング) データと $p' = 257$ の評価 (テスト) データを作成した.

実験方法 2 の結果, 10,473 個のキーワード (属性) が残った. 方法 3 の TF-IDF により,$k = 250$ 個に絞り込む. そのうちの上位 20 位を表 6 に示す.

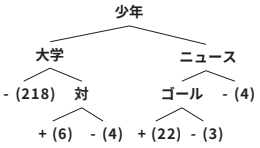

図 3: 2 値クラス決定木:T_A $(min = 3, cf = 7, S(T_A) = 11)$

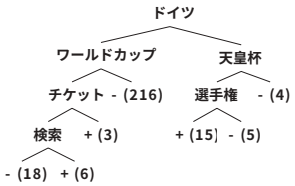

図 4: 2 値クラス決定木:T_B $(min = 3, cf = 7, S(T_B) = 13)$

図 3 と図 4 に, クラス A(アマチュア) と B(イベント) を識別する 2 値木 T_A と T_B を表す. ただし, 最小クラス $min = 3$, 信頼度 $cf = 7$ で簡単化をした結果である. 同様に多値木の結果を図 4 に示す.

表 3: 識別結果 (ランダムに 3 回行った試行の平均)

			標準偏差
2 値	合計サイズ $\sum S(T_i)$	404	5.77
	平均サイズ $S(T_i)$	44.94	5.61
	合計誤差 $\sum E(T_i)$	183	20.3
	平均誤差 $E(T_i)$	20.37	13.3
多値	合計サイズ S	203	6.9
	合計誤差 E	141	3.1
	平均誤差 $E(T_i)$	15.6	12.3

表 4: 多値決定木 誤差表 (ページ数)

真 \ 結果	A	B	C	D	E	F	G	H	I
A	34	1	1		1		2	1	4
B	4	20	1			1		1	3
C			5				1	1	1
D	4			1					1
E	1				9		1	1	6
F	1	2				7		1	2
G	1	2					13		2
H		1			1			6	
I	9	3			6	3			92

方法 5 の多値木評価結果を, 表 4 に示す. ここで, 対角線上に正しく識別されたページ数を, それ以外が誤ったページ数を表している. 方法 6 の 2 値木の評価結果の一部を表 5 に示す. 比較の為, 多値木 T_M を用いて単一クラスを識別した結果も与えている. 例えば,$T_M(x) =$“A” の時は “+”, $T_M(x) \neq$“A” の時は “−” を割り当てて, 2 値木をみなした時の誤差表を示している.

方法 7 に従って, 3 回サンプリングを繰り返して, 求めた平均値を表 3 に整理した. ただし, 2 値木のサイズと誤差は,9 クラスの総和, $S(T_A) + \cdots + S(T_I)$ と $E(T_A) + \cdots + E(T_I)$ である. 信頼度 cf と最小事例数 min によって, 変る枝刈りの影響を図 2 で図示した. ここで,$cf = 0$ から 100 まで,$min = 1, 2, 3$ 変化させている.

クラス数 c によって, 誤差とサイズがどのように変化するか明らかにするために, 同じ実験データを用いて複

表 5: 2 値木と多値木の誤差表 (一部)

		2 値		多値		
真値 \ 識別結果		+	−	+	−	
クラス	+	30	14	34	10	44
A	−	19	194	20	193	213
	計	49	208	54	203	257
クラス	+	19	11	20	10	30
B	−	20	207	9	218	227
	計	39	218	29	228	257

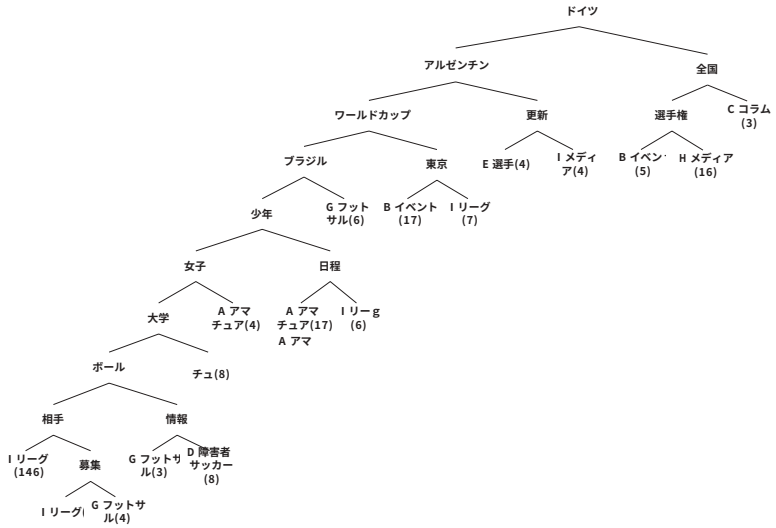

図 5: 多値クラス木 ($T_{min} = 3, cf = 7, S(T_M) = 33$)

数クラスを統合してクラス数 c を変化させて, 実験を繰り返した. その結果を図 6(サイズの比較) と図 7(誤差比較) に各々示す. $c = 9$ の時の値が, これまでの実験結果に相当している.

表 6: TFIDF 値 上位 20 キーワード

順位	キーワード	TFIDF	順位	キーワード	TFIDF
1	ゴシック	.153	11	サイト	.121
2	月	.139	12	試合	.119
3	大会	.138	13	情報	.117
4	代表	.135	14	応援	.1158
5	日本	.133	15	チーム	.1152
6	選手	.132	16	サッカー	.113
7	更新	.131	17	ページ	.1118
8	年	.127	18	掲示板	.1117
9	結果	.126	19	リンク	.110
10	リーグ	.125	20	横浜	.106

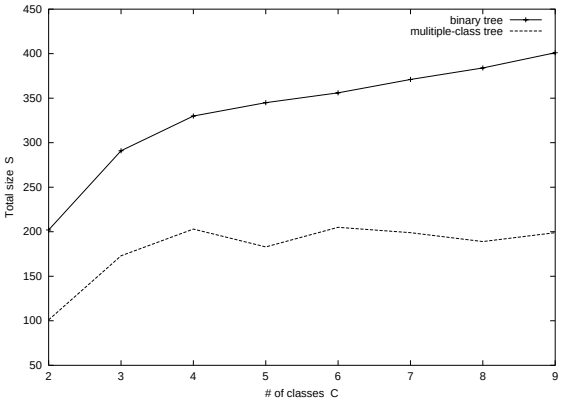

図 6: 2 値-多値 合計サイズ比較

り, 最適な枝刈りをする際には min を変化させるよりも cf を変化させる方が効果的である. 全てのクラスを出力するためのサイズと誤差率の最小値は, $m = 3$, $c = 7$ であった.

4 考察

4.1 最適切な枝刈りについて

最も信頼度を上げる (m=1,cf=100) と, 木のサイズは 189 で, 誤差率は 27.2%だった (図 2). しかし, 図 2 より ,cf の増加に対して, サイズは増大し, 誤差は減少している. しかし, 誤差よりもサイズの変化が著しい. 例えば, m_1 について,$cf = 1$ と 100 の時とのサイズの差は, 152 も変化しているのに対し, 誤差率は 15.2%であった. また, 最小事例誤差 m_1, m_2, m_3 では, サイズの大きさの差が目立つ. それに対して, 誤差の差は 10%程度である. 以上よ

4.2 2 値木と多値木の比較

表 3 を見てみる. 2 値木での合計サイズは 404 に対し, 多値木での合計サイズは 203 と約半分の値であった. 平均サイズでは, 2 値木は約 45 であるが, これは 1 クラスの場合であり, 9 クラスでは,$45 \times 9 = 405$ になり, 平均サイズでも, 多値木の 2 倍近い値を取る. 図 6 では, クラス数 c を変化させて, 木のサイズを観察している. 2 値での決定木では,c が増加するにつれて単調増加している

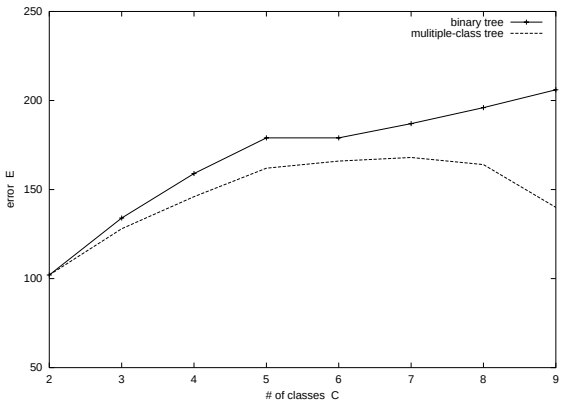

図 7: 2 値-多値 合計誤差比較

図 8: 関東サッカーリーグ

ことがわかる. しかし, 多値木では c が増加しても, 増加は緩やかである.

誤差についてみても, 多値の方が少ない. 図 7 でも, 常に多値の方が少ない値を取っている. 2 値と多値の誤差の原因は決定木自体にあると考えられる.

図 8 は, クラス A(アマチュア) の決定木 (2 値, 多値の両方共) で識別に失敗したページ [7] である. これは明らかに,「アマチュア」カテゴリに属するページだが, T_A は '$-$'(否定),T_M は,I「リーグ (J リーグ情報等)」のカテゴリであるとの識別結果であった. 失敗の原因について, 図 3 の T_A から考察しよう. T_A では, まず「少年」で分類している. ところが, このページは,「少年」は含んでおらず, また次の「大学」も含んでいないため,"$-$" として識別された. アマチュアクラスのほとんどが「少年」や「大学」を対象としている中, 本ページは例外的である.

図 5 の多値木では, 「ドイツ:無」→「アルゼンチン:無」→「ワールドカップ:有」→「東京:有」より,「リーグ」クラスとして識別された. これは, ページの中の一分に,「今シーズンはワールドカップの影響で $\cdots$」と書かれていたためである. このページは「ワールドカップ」には全く関係のないページだが, このキーワードが原因で「リーグ」クラスに分類された.

また, 分類に失敗されるようなページは, トップページに情報が少ない. トップページには最小限サッカーのページであることを示しており, 内容についてはリンクを辿らなければ見ることができないような形式になっている. そのため, そのトップページからは有益な情報が得られず, 誤差を上げる原因にもなっていると考えられる.

図 8 も含め, 識別に失敗するページには, 多ページへリンクを貼っているページが多い. yahoo では, 人間の主観により, Web ページの識別を行っているため, リンク情報が多いページでも的確に分類してしまう.

5 結論

誤差は約 5%多値木の方がよく, サイズについては多値木は 2 値木のサイズの 50%削減できることがわかった. ページ分類問題には, 多値木の方が有効である. しかし, 先ほどの図 8 のように, 現実の多くのページには例外や変則的なページが少なくないため, 本稿のシステムではまだ自動分類には至らない. トップページには有益な情報が少ないという観察から, リンクを辿り, その Web ページについて, より正確な情報を知る必要がある. クラスと誤差の関係を明らかにする事を今後の課題とする.

参考文献

[1] 高崎, 米山, 菊池, 中西, Web ページの分類におけるキーワードの抽出について, 第 7 回気持のワークショップ, pp.9-14, 11 月 2002.

[2] K. Mori, M. Umano, H. Satoh and Y. Uno, "Fuzzy C4.5 for Generating Fuzzy Decision Trees and Its Improvement", *Asian Fuzzy Systems Symposium*,, Tsukuba, Japan, pp.881-884, May 2000.

[3] Quinlan,J. R., *C4.5:Programs for Machine Learning*, Moegan Kaufmann Publishers, 1993.

[4] 菊池浩明, 中西祥八郎, 平均深さの観点から見た決定木学習アルゴリズムの評価, FSS, FE5-3, pp.519-522, 6 月 1999.

[5] 櫻井茂明, 酢山明弘, 布目光生, キー概念辞書を利用しない構造抽出ルールの学習, FSS, pp.81-84,8 月 2002.

[6] http://www.yahoo.co.jp/ (2002 年 12 月参照)

[7] http://www14.u-page.so-net.ne.jp/kb3/ksl/ (2002 年 12 月参照)

[8] http://www.google.com/press/zeitgeist.html/ (2002 年 7 月参照)