人工知能

氏名：津波古正輝　(075739)
桃原岳史　(075741)
橋本達史　(075750)
提出日：12月12日(金曜日)

(1)C4.5プログラムにより、講義の例題を実行し、結果を得る。

sample.dataより決定木を作成。

```
---実行結果より抜粋---
Decision Tree:
eye-color = brown: - (4.0)
eye-color = blue:
|    hair-color = blond: + (2.0)
|    hair-color = black: - (2.0)
|    hair-color = red: + (1.0)
---end---
```

図：sample.dataから決定木の作成

出力された決定木より、目の色が青で髪がブロンドの人物が+になりやすい。
逆に目の色が茶色の人は-になりやすい。

作成された決定木をもとにsample.testの内容を吟味する。

```
---実行結果より抜粋---
            Before Pruning                 After Pruning
           ---------------     ------------------------
           Size         Errors     Size         Errors     Estimate
           6       2(100.0%)          6       2(100.0%)       (43.6%)     <<
             (a)    (b)          <-classified as
             ----  ----
                    1          (a): class +
             1                 (b): class -
---end---
```

図：testでの実験

sumple.testの内容は
low, red, brown, +
low, red, blue, -
であり、2例とも作成された決定木のルールに合致しない内容であった。
そのためerrorの値が高い(100.0%).

(2)自信で分類問題を設計し、
その問題に対する決定木を生成する。

内容

(1)目的：

K-1 GP大会の準決勝以上の試合に出場する選手の属性を用いて決定木を作成し、決勝、準決勝に進出する選手の傾向をさぐる。

(2)データの設計：

今回の課題では、

格闘スタイル：kickboxing、karate、muyathai、boxing、other

年齢：youngは27歳まで、midは28~36、oldは37歳以降

出身地域：asia、europe、africa、southamerica、other

身長：heightは184までがlow,185~195がmid,196以降はhigh

体重：weightは110までがlight,111以上がheavy

テクニック：punch,kick,multi

この６項目を属性として取り入れた。

準決勝戦まで残る選手を+(決勝戦、又は準決勝戦に進出できる),それ以外は-(準決勝戦に進出できない)とした。

注意：最初は決勝戦に進出できる選手の傾向を探る決定木を作成したが、のちに準決勝に進出できる選手の傾向を探る決定木に変わっている。

(3)データの収集

k-1公式サイトから試合結果を参照した。

http://www.k-1.co.jp/databace

各選手の得意技はwikipedia上の記述を参考にした。

http://ja.wikipedia.org/wiki/

(4)どのようにして作成したデータを評価するか。

実際にプラスの要素を持った選手のデータを使用して進出できるかどうか調べる。

2000 年から 2008 年まで 9 年分の WORLD グランプリの WGP 級の試合のデータを集めた。

```
muaythai,young,africa,mid,light,punch,+
muaythai,mid,southamerica,mid,light,kick,+
kickboxing,mid,europe,low,light,punch,+
kickboxing,young,southamerica,mid,heavy,multi,+
kickboxing,mid,europe,mid,heavy,punch,-
kickboxing,mid,europe,mid,light,kick,-
karate,mid,southamerica,mid,light,kick,-
kickboxing,old,europe,mid,heavy,kick,-
```

例：2008 年度のデータ

C4.5 プログラムの実行
必要な部分のみ抜粋する。

```
Read 73 cases (6 attributes) from ../K1/k1.data
style = boxing: - (5.0)
style = other: - (5.0)
style = kickboxing:
|   area = asia: - (3.0/1.0)
|   area = europe: - (27.0/4.0)
|   area = africa: + (3.0/1.0)
|   area = southamerica: - (2.0)
|   area = other: - (1.0)
style = muaythai:
|   weight = heavy: - (3.0)
|   weight = light:
|   |   area = asia: - (1.0)
|   |   area = europe: + (3.0/1.0)
|   |   area = africa: + (2.0/1.0)
|   |   area = southamerica: - (3.0/1.0)
|   |   area = other: - (0.0)
style = karate:
|   height = low: - (1.0)
|   height = high: + (3.0)
```

```
|   height = mid:
|   |   area = asia: - (5.0/2.0)
|   |   area = europe: - (1.0)
|   |   area = africa: + (1.0)
|   |   area = southamerica: - (4.0/1.0)
|   |   area = other: - (0.0)

Simplified Decision Tree:
 - (73.0/21.2)
Evaluation on training data (73 items):
         Before Pruning           After Pruning
        ----------------   --------------------------
        Size      Errors   Size      Errors   Estimate
         26      12(16.4%)     1      18(24.7%)    (29.0%)   <<
```

図：決定木の作成

この決定木は、『決勝まで残る選手の傾向』を調べようと思ったものだ。

注目したいのが木のサイズを示す Size の項目である。最初は 26 であったのに、After Pruning を行った後が、Size が 1 となっている。Simplified Decision Tree を見てみると、『-』ただ 1 つであった。つまり、どのような要素を持った選手でもマイナスと判断されてしまう。つまり、どんな選手でも決勝戦に進出できない、という結果なのである。これはおかしい結論である。実際に打ち込んだデータは決勝戦に進出したプラスの要素を持った選手のデータもある。しかし、その選手と同じ要素を持った選手のデータを決定木でプラスかマイナスかを判断するとマイナスと結論がだされてしまう。

簡略化された決定木がマイナスしか結論をださないというのは、理由がある。プラスという結論を持った選手が少なすぎて、C4.5 が決勝戦に進出する為の要素を確定できず、平均的に見てマイナスを結論として決定してしまう。なのでマイナスしか結論がでない。

本当に結論がマイナスしかでないのかを調べてみる。決定木を作る上で参考にした全データをテストデータに入力し、テストデータでの実験行ってみると、

```
Evaluation on training data (73 items):
         Before Pruning           After Pruning
        ----------------   --------------------------
        Size      Errors   Size      Errors   Estimate
          26   12(16.4%)      1   18(24.7%)    (29.0%)   <<
Evaluation on test data (23 items):
         Before Pruning           After Pruning
        ----------------   --------------------------
        Size      Errors   Size      Errors   Estimate
          26    4(17.4%)      1    7(30.4%)    (29.0%)   <<

          (a)  (b)      <-classified as
         ----  ----
                 7      (a): class +
                16      (b): class -
```

図：test データでの実験

となった。一番下の classified as を見てみると、プラスに出力されているものが１つもない。

　プラスの要素を持つというのは、決勝戦に進んだ選手ということであるが、決勝戦に進む選手というのは数が少なすぎるので、このような結果となった。なので、もっとプラスの要素を持つ選手を増やす事にする。

　ランクを少し下げ、準決勝まで進出する選手にプラスの要素を与えることにする。

準決勝まで進出する選手の要素を探すための決定木を作成する。

```
Read 72 cases (6 attributes) from ../k3/k1.data
style = muaythai: + (11.0/3.0)
style = boxing: - (0.0)
style = other: - (5.0/1.0)
style = kickboxing:
|   area = africa: + (3.0/1.0)
|   area = southamerica: + (4.0/2.0)
|   area = other: + (1.0)
|   area = asia:
|   |   weight = light: - (2.0)
```

```
|   |   weight = heavy: + (2.0)
|   area = europe:
|   |   age = young: + (3.0/1.0)
|   |   age = old: + (4.0/2.0)
|   |   age = mid:
|   |   |   height = low: + (3.0/1.0)
|   |   |   height = high: - (0.0)
|   |   |   height = mid:
|   |   |   |   weight = light: - (11.0/1.0)
|   |   |   |   weight = heavy:
|   |   |   |   |   technique = punch: - (8.0/3.0)
|   |   |   |   |   technique = kick: + (2.0)
|   |   |   |   |   technique = multi: + (0.0)
style = karate:
|   weight = light: - (9.0/3.0)
|   weight = heavy: + (4.0)

Simplified Decision Tree:
style = kickboxing: - (43.0/21.8)
style = muaythai: + (11.0/4.6)
style = boxing: - (0.0)
style = other: - (5.0/2.3)
style = karate:
|   weight = light: - (9.0/4.5)
|   weight = heavy: + (4.0/1.2)

Evaluation on training data (72 items):
         Before Pruning           After Pruning
        ----------------   --------------------------
        Size      Errors   Size      Errors   Estimate
         26     18(25.0%)    8      26(36.1%)    (47.8%)   <<
```

図：準決勝に進出する選手の傾向を調べるための決定木

Before Pruning、After Pruning の Size を見ると、26→8 となっている。簡略化は行われたが、前回のように全てマイナスという決定木にはなっていない。

（以下ページの実行プログラムはすべて準決勝に進出する要員を調べたものである）

この決定木をテストデータ(最強と思われる人間の設定データと試合でよく負ける実在の選手のデータ)を使用して妥当性を示す。

テストデータ

```
muaythai,young,africa,high,heavy,punch,+
other,old,africa,mid,light,multi,-
other,old,europe,high,heavy,multi,-
```

(最強の人間はムエタイ使いで 27 歳以下、アフリカ生まれで身長、体重ともに 2m、120kg。得意技はパンチ)

実行

```
          (a)  (b)      <-classified as
         ----  ----
           1           (a): class +
                2      (b): class -
```

適正に判断された。

C4.5 の rules

```
Read 72 cases (6 attributes) from ../k3/k1
Rule 1:
        area = asia
        weight = light
        ->  class -  [70.7%]
Rule 5:
        age = mid
        area = europe
        height = mid
        ->  class -  [61.7%]
Rule 11:
        style = other
        ->  class -  [54.6%]
Rule 8:
        style = muaythai
        ->  class +  [57.9%]
Rule 6:
        weight = heavy
```

```
        ->  class +  [52.8%]
Default class: +
Evaluation on training data (72 items):
Rule  Size  Error  Used  Wrong            Advantage
----  ----  -----  ----  -----            ---------
   1     2  29.3%     4      0 (0.0%)        4 (4|0)    -
   5     3  38.3%    21      6 (28.6%)       9 (15|6)   -
  11     1  45.4%     5      1 (20.0%)       3 (4|1)    -
   8     1  42.1%    11      3 (27.3%)       0 (0|0)    +
   6     1  47.2%    12      3 (25.0%)       0 (0|0)    +
Tested 72, errors 21 (29.2%)   <<
          (a)  (b)      <-classified as
         ---- ----
           28    7      (a): class +
           14   23      (b): class -
```

図：rule の作成

Rule のカッコの中に表示されるのは信頼率である。
最後に出力される表は、Rule を適用して決定木を作成する際に使用したデータをクラス分けした時の表である。

表の見方：

左上：実際のデータでプラス要素を持っていて、C4.5 がプラスと判断した数
右上：実際のデータでプラス要素を持っていて、C4.5 がマイナスと判断した数
左下：実際のデータでマイナス要素を持っていて、C4.5 がプラスと判断した数
右下：実際のデータでマイナス要素を持っていて、C4.5 がマイナスと判断した数

consult で決勝戦に過去進出した人のデータを入力してどのような結果か調べる。

R・B さん

```
style: muaythai
Decision:
        +  CF = 0.73  [ 0.58 - 1.00 ]
```

S・S さん

```
style [ muaythai ]: karate
weight: heavy
Decision:
        +  CF = 1.00  [ 0.71 - 1.00 ]
```

M さん

```
style [ karate ]: karate
weight [ heavy ]: light
Decision:
        -   CF = 0.67  [ 0.50 - 1.00 ]
```

M・H さん

```
style [ karate ]: kickboxing
Decision:
        -  CF = 0.56  [ 0.49 - 1.00 ]
```

以上の 4 人は決勝戦に進出した人達だが、マイナスと判断される人もいる。
信頼率が高くないので、決勝に進むかもしれない人、と考えてもよいだろう。このように、実際にはプラス要素を持っているが、マイナスと判断される場合もあった。

考察：
　最初にどのような選手がプラスの要素を持つか調べていた際、決定木を作成するデータの中にプラスの要素を持った選手が少なすぎると、全体的に多い要素をとりあえず出力という決定木になってしまった(決勝戦に進出する選手が持つ要素を調べる時の決定木)。つまり、調べたいもの（プラスを持っている選手）が少なすぎると、悲観的な結論しか出力してくれない。よって、ある程度数が多いテーマに設定しないといけない。

まとめ：
　決定木を作成するにあたってのデータが確実なものばかりであれば、確実な結果がでるが、今回調べた K-1 のように不確実なデータがもとになる決定木の結論は確実なものではない。しかし、計算によって信頼率(当る確率)が判明するので何かを決める時の参考にはなる。

参考文献：

k-1公式サイト
http://www.k-1.co.jp/databace

http://ja.wikipedia.org/wiki/