人口知能

-中間レポート-

055702B　　池野谷克俊

提出日 2007 年 1 月 9 日 火曜日

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。

- 決定木作成

```
[j05002@Src]% c4.5 -f ../Sample/sample

C4.5 [release 8] decision tree generator        Tue Dec 26 01:38:24 2006
----------------------------------------

    Options:
        File stem <../Sample/sample>

Read 8 cases (3 attributes) from ../Sample/sample.data

Decision Tree:

eye-color = brown: - (3.0)
eye-color = blue:
|   hair-color = blond: + (2.0)
|   hair-color = black: - (2.0)
|   hair-color = red: + (1.0)


Tree saved


Evaluation on training data (8 items):

         Before Pruning           After Pruning
        ----------------   ---------------------------
        Size      Errors   Size      Errors   Estimate

           6    0( 0.0%)      6    0( 0.0%)    (48.3%)   <<
```

c4.5 によって作成された決定木は自分で問題を解いて作成した決定木と一致した。

- sample.test による評価

```
[j05002@Src]% c4.5 -f ../Sample/sample -u

C4.5 [release 8] decision tree generator        Tue Dec 26 01:42:23 2006
----------------------------------------

    Options:
        File stem <../Sample/sample>
        Trees evaluated on unseen cases

Read 8 cases (3 attributes) from ../Sample/sample.data

Decision Tree:

eye-color = brown: - (3.0)
eye-color = blue:
|   hair-color = blond: + (2.0)
|   hair-color = black: - (2.0)
|   hair-color = red: + (1.0)


Tree saved
```

```
Evaluation on training data (8 items):

         Before Pruning           After Pruning
        ----------------   --------------------------
        Size      Errors   Size      Errors   Estimate

           6    0( 0.0%)      6    0( 0.0%)    (48.3%)   <<

Evaluation on test data (2 items):

         Before Pruning           After Pruning
        ----------------   --------------------------
        Size      Errors   Size      Errors   Estimate

           6    2(100.0%)      6    2(100.0%)    (48.3%)   <<


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

sample.test による評価を行った結果、sample.test にある二つの事例の場合、決定木は正しい予測を行えなかった。本来+と分類されるべき要素が − と分類され、本来 − と分類されるべき要素が + と分類されている。

2 自身で分類問題を設計し、その問題に対する決定木を作成する。

2.1 問題意図

J リーグ (J1) において、リーグ順位が上位 (5 位以内) であるチームの特徴を分析する。

2.2 属性及び属性値の設定

属性として

- シュート数
- 被シュート数
- フリーキック数
- コーナーキック数
- 警告数
- 退場数

を使用する。

属性値は、各属性の平均を計算し、その値に基づき”多い (many)”, ”ふつう (normal)”,”少ない (little)”とする。シュート数、被シュート数、フリーキック数、コーナーキック数に関しては平均より 20 以上高ければ多い、20 以上小さければ少ない、その間をふつうとした。警告数、退場数に関しては、平均より 3 以上高ければ多い、3 以上小さければ少ない、その間をふつうとした。

クラスは上位５チームを”＋”とし、それ以外を”－”とした。

2.3 C4.5 プログラムの作成

- 決定木の作成

```
[j05002@Src]% c4.5 -f ../soccer/soccer

C4.5 [release 8] decision tree generator        Tue Dec 26 17:08:20 2006
----------------------------------------

    Options:
        File stem <../soccer/soccer>

Read 18 cases (6 attributes) from ../soccer/soccer.data

Decision Tree:

shoot = normal: - (3.0/1.0)
shoot = little: - (9.0)
shoot = many:
|   re-shoot = many: - (2.0)
|   re-shoot = normal: + (0.0)
|   re-shoot = little: + (4.0)


Tree saved


Evaluation on training data (18 items):

         Before Pruning           After Pruning
        ----------------   ---------------------------
        Size      Errors   Size      Errors   Estimate

           7    1( 5.6%)      7    1( 5.6%)    (30.6%)   <<
```

2.4 得られた決定木の分析

得られた決定木から、シュート数が多く、被シュート数が普通以下のチームが上位になるということが分かった。しかしシュート数は普通だがクラスが＋の事例が１つ存在しているのでエラーが１となっている。
フリーキックやコーナーキックが絡んでくるかと予想したが、シュートとい

う点数に一番関わる属性が重要と判断されており、とても分かりやすい結果となっている。

2.5 テストデータによる評価

```
[j05002@Src]% c4.5 -f ../soccer/soccer -u

C4.5 [release 8] decision tree generator        Tue Dec 26 17:15:28 2006
----------------------------------------

    Options:
        File stem <../soccer/soccer>
        Trees evaluated on unseen cases

Read 18 cases (6 attributes) from ../soccer/soccer.data

Decision Tree:

shoot = normal: - (3.0/1.0)
shoot = little: - (9.0)
shoot = many:
|   re-shoot = many: - (2.0)
|   re-shoot = normal: + (0.0)
|   re-shoot = little: + (4.0)


Tree saved


Evaluation on training data (18 items):

         Before Pruning           After Pruning
        ----------------   ---------------------------
        Size      Errors   Size      Errors   Estimate

           7    1( 5.6%)      7    1( 5.6%)    (30.6%)   <<

ERROR:  case 11's value of 'little' for attribute taijou is illegal

ERROR:  case 30's value of 'little' for attribute taijou is illegal

Evaluation on test data (34 items):

         Before Pruning           After Pruning
        ----------------   ---------------------------
        Size      Errors   Size      Errors   Estimate

           7    5(14.7%)      7    5(14.7%)    (30.6%)   <<


          (a)  (b)      <-classified as
         ----  ----
            8    2      (a): class +
            3   21      (b): class -
```

2004 年 (16 チーム),2005 年 (18 チーム) のテストデータを用いてルールを評価してみたところ、本来 + となるべきものが − となっているのが２つあり、その逆が３つあるので、正しく判定できているのは３４個中２９個となり、約８割５分の確率で正しく判定できると分かった。結果としては、精度はある程度良い。

2.6　考察

　決定木から分かったようにやはり、サッカーというスポーツにおいては得点に一番絡んでくる、シュートという行動が重要なようだ。
確かに、シュートをたくさん打ち、相手にシュートさせないようにするということが勝利に繋がるということは、サッカーをあまり知らない人でも予想することができるだろう。
しかし、フリーキックやコーナーキックというのはシュートに繋がる行動であり、世界の強豪チームはこのようなセットプレーから点数を決めるというのが多い。
今回のデータでは、シュートとコーナーキックやフリーキックの関係性などが考慮されることはないので、この二つの属性は重要ではないと判断されているが、本来は重要な属性であると考えられる。もし、このルールが正しいと判断するならば、Ｊリーグがセットプレーからの得点が少ないという特徴を持っている、ということになる。