2022年度 知識工学 レポート課題

氏名:
学生番号:

2022年5月29日

1 課題1

1.1 課題

決定木学習モデルを download し，windows または linux にインストールして付属のデモファイルについて解説をする．ない場合は，簡単な例を作成して動作説明をする（解説 blog などで利用されているデータを参考にして良い）

さらに動かした結果を付与し説明を加えよ．いくつかのデモが付与されているが 1 つで良い．

1.2 手法

RuleQuest Research[1] の配布する，See5/C5.0 のデモプログラムを実行する．

付属のデモファイル `anneal.{data,names,test}` を使用した．

1.3 デモファイル

1.3.1 anneal.names

`anneal.names` はスキーマファイルである．後述する `anneal.data`，`anneal.test` のクラス，各属性の型を定義する．

anneal.names

```
1,2,3,5,U.

family:         discrete 12.|GB,GK,GS,TN,ZA,ZF,ZH,ZM,ZS.
product-type:   C, H, G.
steel:          R,A,U,K,M,S,W,V.
carbon:         continuous.
hardness:       continuous.
```

上記が `anneal.names` の冒頭の抜粋である．

1 行目はクラスを定義する．与えるデータが 1, 2, 3, 5, U の 5 つのクラスに分類される．

3 行目以降は各属性について定義する．3 行目は `family` という属性値について定義する．`family` は**最大で** 12 種類の離散値を取ることができる．`|` 以降の記述はコメントであることに注意されたい．

4,5 行目はそれぞれ `product-type`，`steel` という属性値について定義する．これらはそれぞれ C, H, G と R, A, U, K, M, S, W, V のうちのいずれかの値を取ると明示的に指定している．

6,7 行目はそれぞれ `carbon`，`hardness` という属性値について定義する．これらは連続値（実数値）を取ることができる．

1.3.2 anneal.data, anneal.test

anneal.data は学習データ，anneal.test はテストデータであり，同じ形式を取る．

anneal.data

```
1  TN,C,A,00,00,N/A,N/A,3,000,N,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,C,N/A,N/A,N/A,N/A,N/A,N/A,
2      N/A,N/A,N/A,N/A,N/A,N/A,SHEET,0.600,1220.0,0761,N/A,0000,N/A,5
3  N/A,C,A,00,00,N/A,S,2,000,N/A,N/A,E,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,
4      N/A,N/A,N/A,N/A,N/A,N/A,SHEET,0.699,0610.0,0762,N,0000,N/A,3
5  TN,C,N/A,00,00,N/A,A,1,000,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,Y,N/A,
6      N/A,N/A,N/A,N/A,N/A,N/A,SHEET,0.500,0609.9,0612,N/A,0000,N/A,5
7  N/A,C,A,00,60,T,N/A,N/A,000,N/A,N/A,G,N/A,N/A,N/A,N/A,B,Y,N/A,N/A,N/A,Y,N/A,N/A,N/A,N/A,
8      N/A,N/A,N/A,N/A,N/A,N/A,SHEET,0.800,0356.1,0762,N/A,0000,N/A,3
9  ZS,C,A,00,00,N/A,S,3,000,N,N/A,E,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,
10     N/A,N/A,N/A,N/A,N/A,COIL,1.000,0075.0,0000,N/A,0000,N/A,3
```

上記が anneal.data の冒頭の抜粋である（適当に改行を加えた）．

各行が学習データであり，コンマで区切られたデータが並んでいる．1 列目から順に anneal.names に定義した**属性**が順番に並んでいる．そして**末尾にクラス**が与えられる．

1.4 実行結果

実行結果の全体は付録 A に添付している．ここでは一部を説明するに留める．また，詳細については https://www.rulequest.com/see5-unix.html#CLASSIFIERS に記載されている．

11 行目には学習データの要約が記載されている．デモプログラムでは 400 ケースまでの制限があるため，400 ケース 38 属性のデータが読み込まれたことを示している．

13–38 行目には学習結果が決定木として示されている．上から順に，hardness の属性値が 75 より大きいかそれ以下かに分け，大きければクラス U に分類し，それ以下であれば更に分類する．strength の属性値が 375 より大きいかそれ以下かに分け，大きく，かつ steel の属性値が U, K, W, V, M, N/A のいずれかであればクラス 2 に，R, A のどちらかであればクラス 3 に，S であればクラス 1 に分類する．strength の属性値が 375 以下で，かつ family の属性値が TN であればクラス 5 に，ZS であればクラス 3 に，N/A であれば更に分類．と続く．

また，分類クラスの後ろの括弧付きの数字は，(そのクラスに分類されたケース数/分類が誤っていたケース数) を示している．

41–68 行目には学習データに対する決定木の評価が示されている．Size は分類されたインスタンス数が **0 でない** 決定木の葉の個数を，Errors は実際の（与えられた）クラスと誤って分類されたインスタンスの個数及び割合を示している．

50–57 行目は混同行列と呼ばれる行列であり，行が実際のクラス，列が分類されたクラスを表している．すなわち対角成分が正しく分類されたインスタンス数を，非対角成分が誤って分類されたインスタンス数を表す．

59–68 行目は各属性が学習データの分類をするときに使われたインスタンス数の割合を示している．すべてのインスタンスの hardness が使われ，分類が完了しなかった 95% のインスタンスの strength が分類に使われた，といった具合である．

74–89 行目はテストデータに決定木を適用したときの混同行列である．

2 課題 2

2.1 課題

売れているノートパソコンの人気機種の属性を調べる

- Web サイトの人気順

- 上位 1–10 は正例（売れた）
- 下位 11 以降は負例（売れないと仮定）

1. moodle に C5.0 用のファイルを置いたので実行してどういう木になったか分析
2. 元々のファイル notePC_rank.xlsx から属性を 1 つ選び，追加して分析せよ

2.2 課題 2.1

実行結果は付録 B に示す.
12–20 行目の決定木から分類結果を表にまとめると表 1 のようになる.

表 1　ノート PC の売れ方と価格・CPU スコアの関係

<table>
<tr><td>CPU スコア
価格</td><td rowspan="2">≤ 9019</td><td>> 9019
≤ 11550</td><td rowspan="2">> 11550</td></tr>
<tr><td>N/A</td><td colspan="3">負例 (2)</td></tr>
<tr><td>≤ 49784</td><td colspan="3">負例 (2)</td></tr>
<tr><td>> 49784
≤ 68020</td><td rowspan="2">正例 (4)</td><td colspan="2">正例 (3)</td></tr>
<tr><td>> 68020</td><td>負例 (4)</td><td>正例 (3/1)</td></tr>
</table>

ここからは大まかに価格が 49784 円より高いほうが売れ行きがよく，価格が高いが CPU スコアが 9019 から 11550 の間であるとあまり売れないと読み取れる.

2.3 課題 2.2

ディスプレイサイズの属性を加え，`notepc.data`，`notepc.names` を 付録 C のようにした.
このときの C5.0 の実行結果は 8 行目の属性数を除き 2.2 節と同じであった.

参考文献

[1] RuleQuest Research Data Mining Tools, `https://www.rulequest.com/index.html`, 2022 年 5 月 19 日アクセス.

付録 A　anneal の決定木の学習結果

anneal.output

```
1
2  C5.0 [Release 2.11a]          Thu May 19 20:54:42 2022
3  --------------------
4
5      Options:
6          Application `./C50Release2/Data/anneal'
7
8          ** This demonstration version cannot process **
9          ** more than 400 training or test cases.     **
10
11 Read 400 cases (38 attributes) from ./C50Release2/Data/anneal.data
12
13 Decision tree:
```

```
hardness > 75: U (21/1)
hardness <= 75:
:...strength > 375:
    :...steel in {U,K,W,V}: 2 (0)
    :   steel = N/A: 2 (6)
    :   steel in {R,A}: 3 (5)
    :   steel = M: 2 (6)
    :   steel = S: 1 (2)
    strength <= 375:
    :...family = TN: 5 (35)
        family = ZS: 3 (19)
        family = N/A:
        :...enamelability in {3,4,5}: 3 (0)
            enamelability = 1: 3 (3)
            enamelability = 2: 2 (5)
            enamelability = N/A:
            :...surface-quality in {D,E,F,G}: 3 (240)
                surface-quality = N/A:
                :...condition in {A,X}: 3 (0)
                    condition = N/A: 3 (24)
                    condition = S:
                    :...temper_rolling = -: 2 (0)
                        temper_rolling = N/A: 2 (32)
                        temper_rolling = T: 3 (2)


Evaluation on training data (400 cases):

           Decision Tree
         ----------------
         Size      Errors

           13    1( 0.2%)   <<


          (a)   (b)   (c)   (d)   (e)    <-classified as
         ----  ----  ----  ----  ----
            2                            (a): class 1
                 49                      (b): class 2
                      293           1    (c): class 3
                             35          (d): class 5
                                   20    (e): class U


        Attribute usage:

            100%  hardness
             95%  strength
             91%  family
             77%  enamelability
             75%  surface-quality
             15%  condition
              9%  temper_rolling
              5%  steel
```

```
        ** This demonstration version cannot process **
        ** more than 400 training or test cases.     **

Evaluation on test data (400 cases):

            Decision Tree
          ----------------
          Size      Errors

            13    6( 1.5%)   <<


           (a)   (b)   (c)   (d)   (e)    <-classified as
          ----  ----  ----  ----  ----
             2           3                (a): class 1
                  38                      (b): class 2
                       311           1    (c): class 3
                             25           (d): class 5
                        2          18    (e): class U


Time: 0.0 secs
```

付録 B　notepc_s の決定木の学習結果

notepc_s.output

```
C5.0 [Release 2.11a]          Thu May 19 23:35:59 2022
--------------------

    Options:
        Application `notepc_s'

Read 18 cases (5 attributes) from notepc_s.data

Decision tree:

price = N/A: 0 (2)
price <= 49784: 0 (2)
price > 49784:
:...cpu_score <= 9019: 1 (4)
    cpu_score > 9019:
    :...price <= 68020: 1 (3)
        price > 68020:
        :...cpu_score <= 11550: 0 (4)
            cpu_score > 11550: 1 (3/1)


Evaluation on training data (18 cases):

            Decision Tree
          ----------------
          Size      Errors

```

```
             6    1( 5.6%)   <<


           (a)   (b)    <-classified as
          ----  ----
             8     1    (a): class 0
                   9    (b): class 1


        Attribute usage:

            100%  price
             78%  cpu_score


Time: 0.0 secs
```

付録 C　notepc の.data と.names

notepc.data

```
Dell,80282.0,10088,8,1.06,13.3,0
NEC,111183.0,13803,8,2.1,15.6,0
Dell,37980.0,4146,4,1.59,14,0
Lenovo,49588.0,11275,8,1.5,14,0
Lenovo,N/A,16190,8,2.2,17.3,0
Lenovo,71060.0,10088,8,1.7,15.6,0
Dell,96122.0,10521,16,1.714,15.6,0
マウスコンピューター,N/A,7951,8,1.59,15.6,0
Dynabook,127380.0,10521,16,1.94,15.6,0
NEC,89700.0,6132,8,N/A,15.6,1
Dell,56026.0,7149,8,1.83,15.6,1
マウスコンピューター,129800.0,12580,16,2.13,15.6,1
Dell,49980.0,6324,8,1.78,15.6,1
マイクロソフト,80500.0,7951,8,1.11,12.4,1
Dell,64676.0,13851,8,N/A,14,1
Lenovo,56430.0,11275,8,1.45,14,1
Lenovo,75460.0,13803,16,1.5,14,1
Dell,64980.0,10088,8,1.78,15.6,1
```

notepc.names

```
0,1.

maker:        discrete 6.
price:        continuous.
cpu_score:        continuous.
memory:        continuous.
weight:        continuous.
display:        continuous.
```