دانشگاه صنعتی اصفهان

دانشکده برق و کامپیوتر

عنوان

تمرین شماره 1 درس یادگیری ماشین

استاد درس:

دکتر پالهنگ

دانشجو:

داریوش حسن پور

هوش مصنوعی (9308164)

1393

معرفی مجموعه داده

- **نام:** تأیید اعتبار (Credit Approval)
- **منبع:** quinlan@cs.su.oz.au
- **استفاده های قبلی:**
  - "Simplifying decision trees", Int J Man-Machine Studies 27, Dec 1987, pp. 221-234.
  - "C4.5: Programs for Machine Learning", Morgan Kaufmann, Oct 1992
- **توضیحات:**

فایل مربوط به مجموعه داده ای شامل اصلاعاتی راجع به کارت های اعتباری هست. که با توجه به توضیحاتی که در راجع به فایل آمده نام ویژگی برای حفاظت از داده های اصلی تغییر داده شده اند.

- **تعداد نمونه ها :** 690 عدد
- **تعداد ویژگی ها:** 15 عدد (16 تا به همراه ویژگی کلاس)
- **تعداد کلاس ها:** 2 عدد (+ یا -)

- **ویژگی های و محدودیت دامنه ای آنها:**
  - A1: { b, a }
  - A2: پیوسته
  - A3: پیوسته
  - A4: { u, y, l, t }
  - A5: { g, p, gg }
  - A6: { c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff }
  - A7: { v, h, bb, j, n, z, dd, ff, o }
  - A8: پیوسته
  - A9: { t, f }
  - A10: { t, f }
  - A11: پیوسته
  - A12: { t, f }
  - A13: { g, p, s }
  - A14: پیوسته
  - A15: پیوسته
  - A16: { +, - } (ویژگی کلاس)

- **توزیع کلاس ها:**
  - + : 307 (44.5%)
  - - : 383 (55.5%)

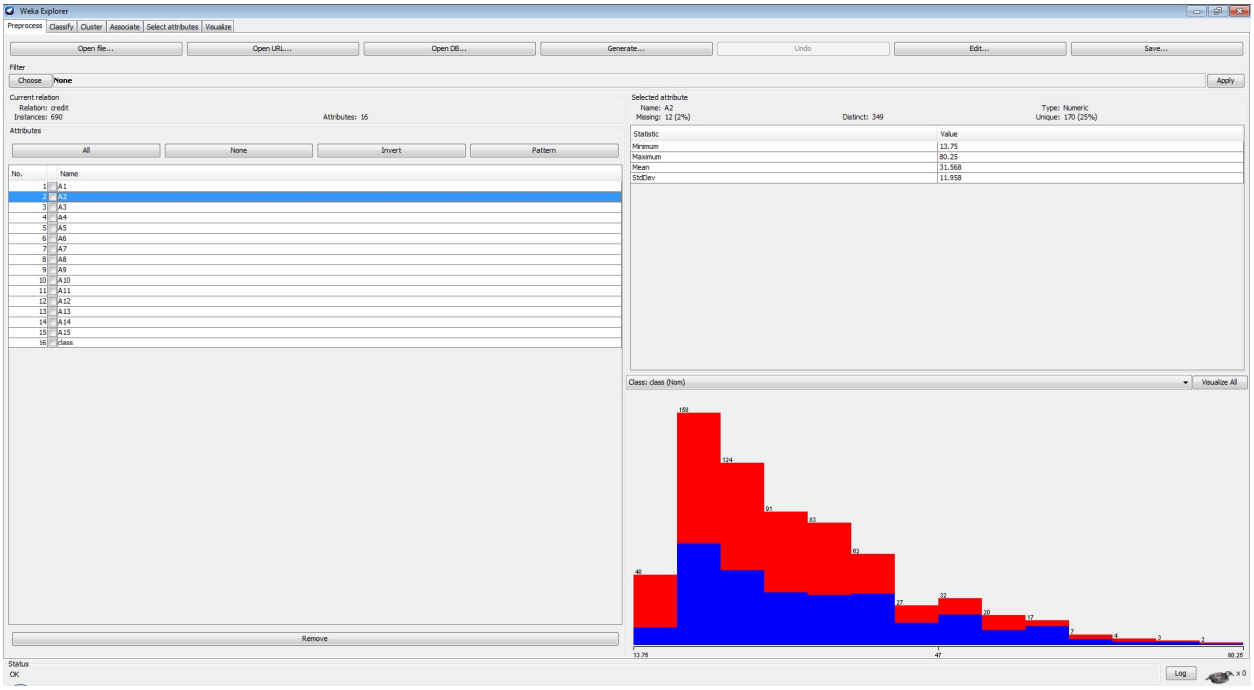
- **آماده سازی داده ها برای استفاده از داده در نرم افزار وکا:**

برای اینکه داده های موجود در UCI را بتوانیم در نرم افزار وکا استفاده کنیم باید بر داده موجود در فایل crx.data گرفته شده از UCI اطلاعات زیر را اضافه کنیم و فایل جدید را با پسوند arff. ذخیره کنیم:

```
@RELATION credit
@ATTRIBUTE A1 {b,a}
@ATTRIBUTE A2 REAL
@ATTRIBUTE A3 REAL
@ATTRIBUTE A4 {u,y,l,t}
@ATTRIBUTE A5 {g,p,gg}
@ATTRIBUTE A6 {c,d,cc,i,j,k,m,r,q,w,x,e,aa,ff}
@ATTRIBUTE A7 {v,h,bb,j,n,z,dd,ff,o}
@ATTRIBUTE A8 REAL
@ATTRIBUTE A9 {t,f}
@ATTRIBUTE A10 {t,f}
@ATTRIBUTE A11 REAL
@ATTRIBUTE A12 {t,f}
@ATTRIBUTE A13 {g,p,s}
@ATTRIBUTE A14 REAL
@ATTRIBUTE A15 REAL
@ATTRIBUTE class {+, -}
@DATA
```

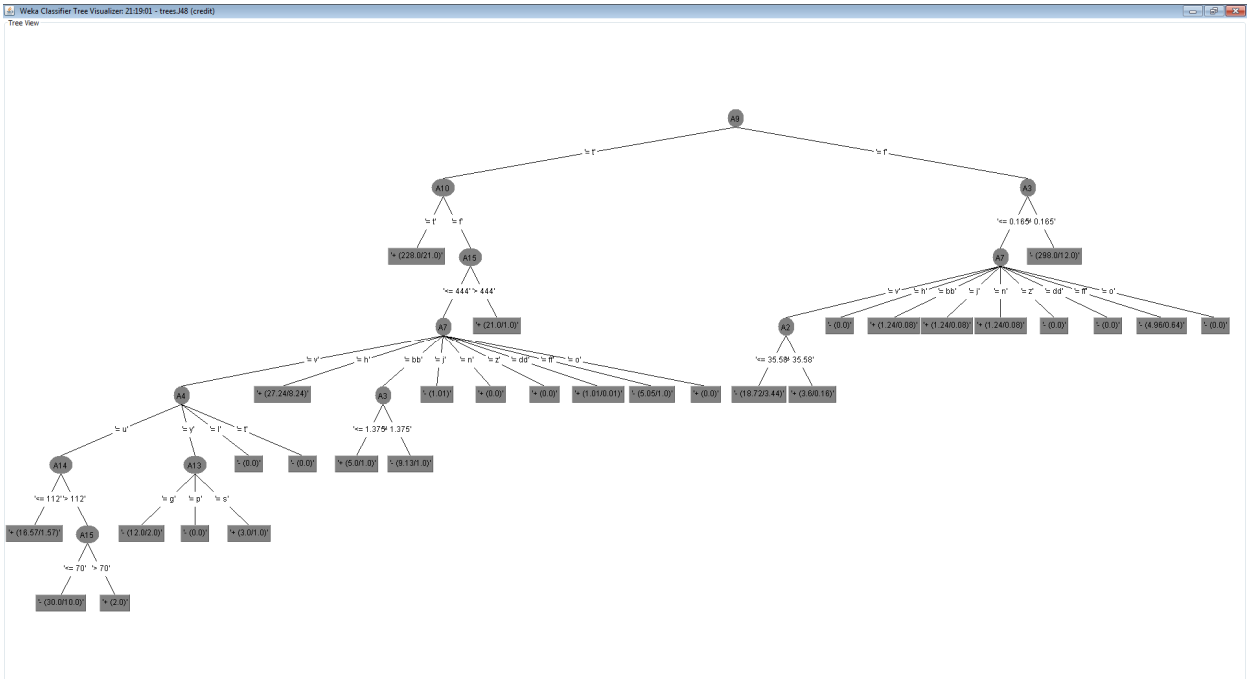
توزیع داده ها در وکا

بعد از اینکه فایل arff. رو در وکا لود کردیم میتوانیم توزیع داده ای مربوط به هر ویژگی رو مشاهده کنیم. به عنوان مثال در ویژگی زیر به سادگی مشاهده میشود که هرکدام از مقادیر ویژگی دارای چه توزیعی است.



ساختن درخت تصمیم با استفاده از الگوریتم C4.5 در وکا

در صورتی که درخت با یک مجموعه آموزشی(70%) و یک مجموعه تست(30%) درخت تصمیم زیر تشکیل می‌شود.



که نتایج آن در مجموعه تست به شرح زیر است:

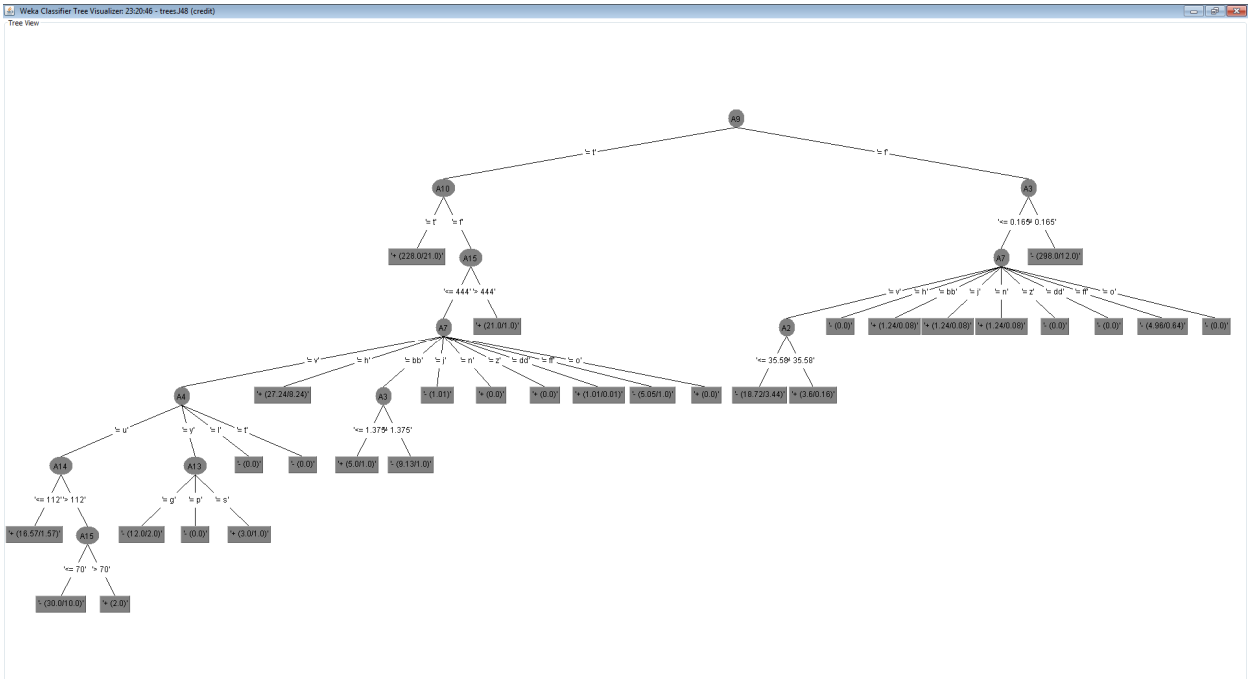
```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         178               85.9903 %
Incorrectly Classified Instances        29               14.0097 %
Kappa statistic                          0.7168
Mean absolute error                      0.1958
Root mean squared error                  0.3288
Relative absolute error                 39.4502 %
Root relative squared error             65.6306 %
Total Number of Instances              207

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.776    0.064      0.916     0.776     0.84       0.901    +
                 0.936    0.224      0.823     0.936     0.876      0.901    -
Weighted Avg.    0.86     0.149      0.867     0.86      0.859      0.901
```

و درخت با تصدیق 3 تایی به شرح زیر است:



که نتایج آن در مجموعه تست به شرح زیر است:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         588               85.2174 %
Incorrectly Classified Instances       102               14.7826 %
Kappa statistic                          0.7003
Mean absolute error                      0.1986
Root mean squared error                  0.3547
Relative absolute error                 40.2113 %
Root relative squared error             71.3827 %
Total Number of Instances              690

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.827    0.128      0.838     0.827     0.833      0.856    +
                 0.872    0.173      0.863     0.872     0.868      0.856    -
Weighted Avg.    0.852    0.153      0.852     0.852     0.852      0.856

=== Confusion Matrix ===

   a   b   <-- classified as
 254  53 |   a = +
  49 334 |   b = -
```

آموزش با استفاده از شبکه عصبی

به تعداد 3 عدد نمونه ویژگی A2 آنها از مقدار اصلی تغییر پیدا کرده و به تعداد 12 عدد از نمونه ها مقدار مجهول ویژگی A2 آنها به طور تصادفی مقدار دهی شده اند؛ با تغییر ویژگی A2 به تعداد 15عدد از داده ها که شامل پر کردن تصادفی داده هایی که مقداری برای این ویژگی نداشتن تغییراتی را در داده ها ایجاد کردیم.

در صورتی که شبکه چند لایه با یک مجموعه آموزشی(70%) و یک مجموعه تست(30%) نتایج زیر را داریم:

```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         166               80.1932 %
Incorrectly Classified Instances        41               19.8068 %
Kappa statistic                          0.6013
Mean absolute error                      0.2051
Root mean squared error                  0.4191
Relative absolute error                 41.3043 %
Root relative squared error             83.6586 %
Total Number of Instances              207

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.755    0.156      0.813     0.755     0.783      0.883    +
                 0.844    0.245      0.793     0.844     0.818      0.883    -
Weighted Avg.    0.802    0.203      0.803     0.802     0.801      0.883

=== Confusion Matrix ===

  a  b   <-- classified as
 74 24 |  a = +
 17 92 |  b = -
```

و شبکه چند لایه با تصدیق 3 تایی به شرح زیر است:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         578               83.7681 %
Incorrectly Classified Instances       112               16.2319 %
Kappa statistic                          0.6707
Mean absolute error                      0.1721
Root mean squared error                  0.3821
Relative absolute error                 34.8443 %
Root relative squared error             76.897  %
Total Number of Instances              690

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.808    0.138      0.824     0.808     0.816      0.89     +
                 0.862    0.192      0.848     0.862     0.855      0.89     -
Weighted Avg.    0.838    0.168      0.837     0.838     0.838      0.89

=== Confusion Matrix ===

   a   b   <-- classified as
 248  59 |   a = +
  53 330 |   b = -
```

مقایسه و نتیجه

برای مقایسه یادگیری های انجام شده، فاکتور های ROC که وکا در خروجی آورده است را در نظر میگیریم که در جدول زیر آمده است.

الگوریتم / روش آموزش	درخت تصمیم	شبکه عصبی
مجموعه آموزشی/تست	0.901	0.883
تصدیق مقاطع 3 تایی	0.856	0.890

همانطور که میبینیم درخت تصمیم با روش آموزشی "مجموعه آموزشی/تست" مقدار ROC بالاتری نسبت به بقیه دارند بنابراین برای یادگیری این مفهوم هدف مناسب است. البته لازم به ذکر است بنابه اینکه در استفاده از شبکه عصبی داده های آموزشی دست کاری شده اند برای همین نمیتوان گفت که شبکه عصبی برای یادگیری این مفهوم مناسب نیست چرا که داده ها دارای نویز بوده اند؛ بنابراین در یک تلاش دیگر با **داده های اصلی بدون دست کاری شده** سعی شده که شبکه عصبی را آموزش داده که در بخش بعدی به آورده شده و مقایسه ای با نتایج درخت تصمیم ساخته شده در قسمت قبل انجام شده است.

شبکه عصبی با داده های بدون دست کاری شده

در صورتی که شبکه چند لایه با یک مجموعه آموزشی(70%) و یک مجموعه تست(30%) نتایج زیر را داریم:

```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         160               77.2947 %
Incorrectly Classified Instances        47               22.7053 %
Kappa statistic                          0.5411
Mean absolute error                      0.2162
Root mean squared error                  0.4355
Relative absolute error                 43.5544 %
Root relative squared error             86.9336 %
Total Number of Instances              207

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.684     0.147      0.807     0.684     0.74       0.866    +
                 0.853     0.316      0.75      0.853     0.798      0.866    -
Weighted Avg.    0.773     0.236      0.777     0.773     0.771      0.866

=== Confusion Matrix ===

  a  b   <-- classified as
 67 31 |  a = +
 16 93 |  b = -
```

و شبکه چند لایه با تصدیق 3 تایی به شرح زیر است:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         568               82.3188 %
Incorrectly Classified Instances       122               17.6812 %
Kappa statistic                          0.6411
Mean absolute error                      0.1934
Root mean squared error                  0.4023
Relative absolute error                 39.1465 %
Root relative squared error             80.9497 %
Total Number of Instances              690

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.788     0.149      0.809     0.788     0.799      0.877    +
                 0.851     0.212      0.834     0.851     0.842      0.877    -
Weighted Avg.    0.823     0.184      0.823     0.823     0.823      0.877

=== Confusion Matrix ===

   a   b   <-- classified as
 242  65 |   a = +
  57 326 |   b = -
```

مقایسه و نتیجه کلی

برای مقایسه یادگیری های انجام شده، فاکتور های ROC که وکا در خروجی آورده است را در نظر میگیریم که در جدول زیر آمده است.

الگوریتم / روش آموزش	درخت تصمیم	شبکه عصبی(با دستکاری داده ای)	شبکه عصبی(بدون دستکاری داده ای)
مجموعه آموزشی/تست	0.901	0.883	0.866
تصدیق مقاطع 3 تایی	0.856	0.890	0.877

باوجود اینکه دادهای بدون دست کاری شده توسط شبکه عصبی آموزش دیده شدند ولی مقدار ROC آن در هر دو روش آموزشی کمتر از زمانی است که داده های شبکه عصبی دست کاری شده اند؛ همانطور که پیشتر گفته شد به تعداد 3 عدد نمونه ویژگی A2 آنها از مقدار اصلی تغییر پیدا کرده و به تعداد 12 عدد از نمونه ها مقدار مجهول ویژگی A2 آنها به طور تصادفی مقدار دهی شده اند؛ که همانطور که نتایج فوق نشان میدهند تغییرات اعمال شده باعث شده اند شبکه عصبی مفهوم را بهتر از زمانی که داده بدون دست کاری استفاده شده اند را یادگرفته است. ولی باز نیز همانطور که میبینیم درخت تصمیم با روش آموزشی "مجموعه آموزشی/تست" مقدار ROC بالاتری نسبت به بقیه دارند بنابراین برای یادگیری این مفهوم هدف مناسب است.