Vilniaus Universitetas

Matematikos ir Informatikos fakultetas

Daugiamačių duomenų vizualizavimas

3 Užduotis

Pagrindinių komponenčių analizė

Laimonas Beniušis
Informatika 1gr.

Turinys

1 Atributai

Duomenų aibė yra pilna (5172 įrašai). Ne visi atributai yra tinkami vizualizavimui, todėl palikti tik šie:

charSeq1	Raidinių žodžių kiekis, kurių ilgis 1 / žodžių kiekis
charSeq2	Raidinių žodžių kiekis, kurių ilgis 2 / žodžių kiekis
charSeq3	Raidinių žodžių kiekis, kurių ilgis 3 / žodžių kiekis
charSeq4	Raidinių žodžių kiekis, kurių ilgis 4 / žodžių kiekis
charSeq5	Raidinių žodžių kiekis, kurių ilgis 5 / žodžių kiekis
charSeq6	Raidinių žodžių kiekis, kurių ilgis 6 / žodžių kiekis
charSeq7	Raidinių žodžių kiekis, kurių ilgis 7 / žodžių kiekis
charSeq8	Raidinių žodžių kiekis, kurių ilgis 8 / žodžių kiekis
charSeq9	Raidinių žodžių kiekis, kurių ilgis 9 / žodžių kiekis
whiteSpaceCount	Tarpinių simbolių (*whitespace*) kiekis / simbolių kiekis
spam	{SPAM,NOTSPAM} E.laiško kategorija

Naudojama WEKA programinė įranga.

2 Sprendimas

2.1 Analizė

Apskaičiuota atributų dispersija pagal įtaką klasei. Galime pastebėti, didžiausią įtaką turi tarpinių simbolių (whitespace) kiekis ir žodžių iš 7 ar 8 simbolių kiekis (žr 1 Pav.).

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 11 spam):
        Information Gain Ranking Filter

Ranked attributes:
 0.1249  10 whiteSpaceCount
 0.1141   7 charSeq7
 0.0708   8 charSeq8
 0.0548   2 charSeq2
 0.0444   3 charSeq3
 0.0408   9 charSeq9
 0.0389   4 charSeq4
 0.0363   5 charSeq5
 0.0347   1 charSeq1
 0.0342   6 charSeq6

Selected attributes: 10,7,8,2,3,9,4,5,1,6 : 10
```

1. Pav: Weka duomenų attributų dispersija

Toliau, pritaikius PrincipalComponents įvertinimo metodą, buvo gauta:

```
Correlation matrix
   1      0.1    0      0.06   0.01   0.02  -0.06  -0.1   -0.07   0.2
   0.1    1      0.05  -0     -0.04  -0.09  -0.13  -0.29  -0.13   0.41
   0      0.05   1      0.25  -0     -0.24  -0      0.04  -0.31   0.19
   0.06  -0      0.25   1      0.12   0.06   0.03   0.01  -0.15  -0.11
   0.01  -0.04  -0      0.12   1      0.15   0.11   0      0.06  -0.24
   0.02  -0.09  -0.24   0.06   0.15   1      0.05   0.13   0.26  -0.36
  -0.06  -0.13  -0      0.03   0.11   0.05   1      0.14   0.11  -0.39
  -0.1   -0.29   0.04   0.01   0      0.13   0.14   1      0.27  -0.43
  -0.07  -0.13  -0.31  -0.15   0.06   0.26   0.11   0.27   1     -0.4
   0.2    0.41   0.19  -0.11  -0.24  -0.36  -0.39  -0.43  -0.4    1
```

2. Pav: Weka duomenų attributų koreliacijos matrica

Taip pat buvo nustatyta kokia procentinė dalis nuo visos dispersijos tenka kiekvienai komponentei (čia atitinka stulpelis *proportion*):

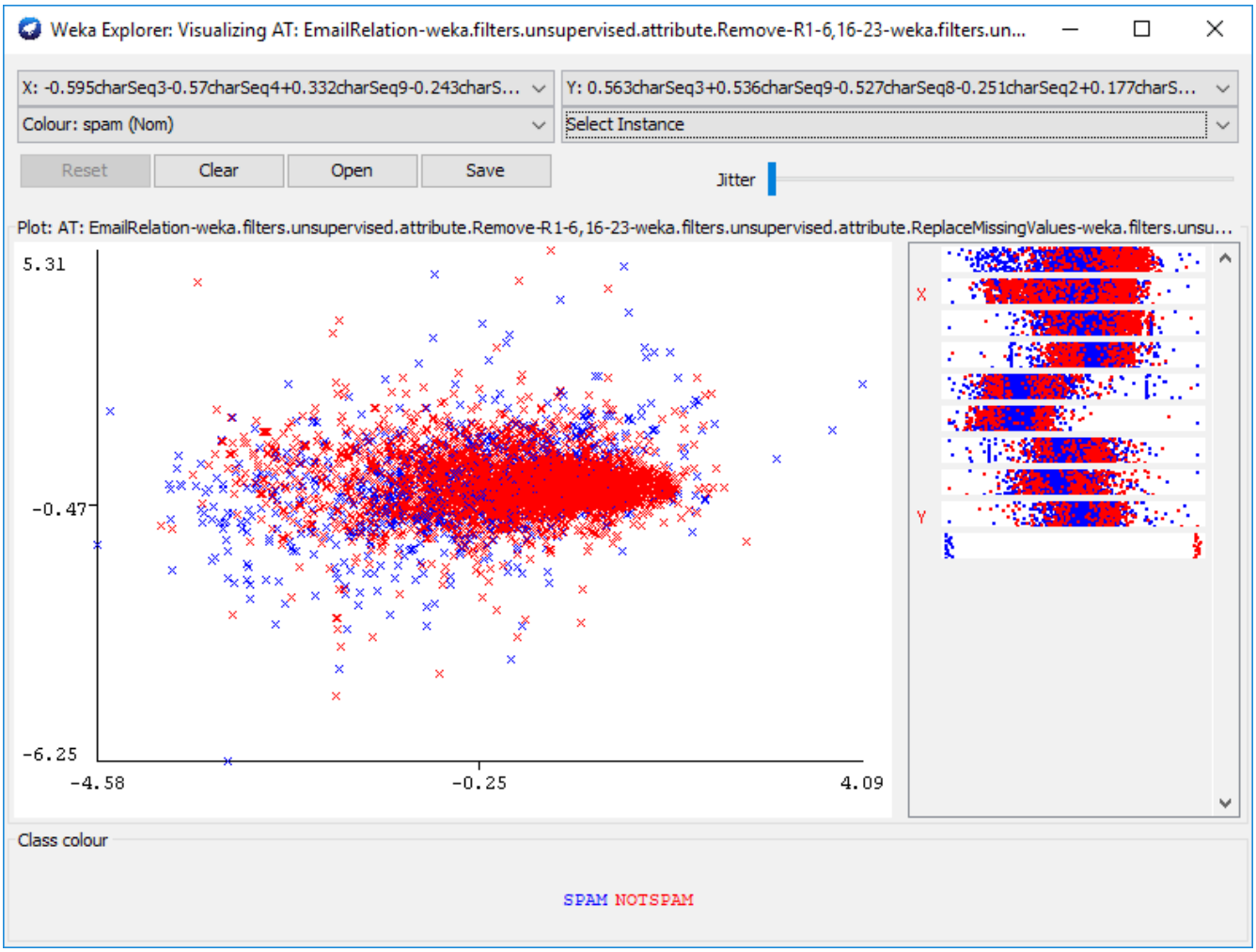
eigenvalue	proportion	cumulative	
2.43944	0.24394	0.24394	0.555whiteSpaceCount-0.396charSeq9-0.37charSeq8+0.338charSeq2-0.327charSeq6...
1.41704	0.1417	0.38565	-0.595charSeq3-0.57charSeq4+0.332charSeq9-0.243charSeq7-0.231charSeq5...
1.20366	0.12037	0.50601	-0.515charSeq5-0.463charSeq6-0.442charSeq1-0.329charSeq4+0.311charSeq8...
0.9595	0.09595	0.60196	0.544charSeq1-0.45charSeq7-0.416charSeq5+0.401charSeq8-0.25charSeq2...
0.90602	0.0906	0.69257	0.661charSeq1+0.644charSeq7-0.273charSeq6-0.257charSeq4-0.068charSeq5...
0.8047	0.08047	0.77304	-0.653charSeq2+0.499charSeq5-0.335charSeq7-0.291charSeq4-0.222charSeq9...
0.76527	0.07653	0.84956	0.454charSeq8+0.44 charSeq5+0.401charSeq2+0.379charSeq9+0.338charSeq3...
0.63647	0.06365	0.91321	0.626charSeq6-0.5charSeq4-0.462charSeq9+0.319charSeq3+0.165charSeq8...
0.55107	0.05511	0.96832	0.563charSeq3+0.536charSeq9-0.527charSeq8-0.251charSeq2+0.177charSeq6...

3. Pav: Weka duomenų attributų proporcijos

2.2 Vizualizacija

4. Pav: Pagridinės komponentės 1-10

5. Pav: Pagrindinės komponentės 1-2



6. Pav: Pagrindinės komponentės 2-10

3 Išvados

Galima pastebėti, kad brukalo tipas grafuose atvaizduojamas kairėje pusėje (žr. 4 Pav., 5 Pav.), tačiau nevisada (žr. Pav. 6). Taip pat verta paminėti, kad didžioji dalis brukalo laiškų atvaizduojami ten pat kur ir paprasti laiškai, todėl tokio principo filtras nebūtų labai geras, visgi tokia analizė parodo, kad įmanoma dalį laiškų atskirti vien tik pagrindinių komponenčių būdu.