

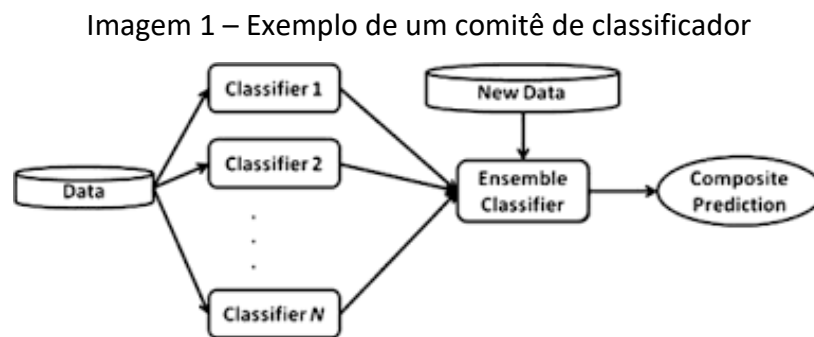
Relatório da disciplina de datawarehouse e datamining

Trabalho sobre comitê de classificadores.

Rodrigo Rafael da Maceno de Souza

Comitê de classificadores

Um comitê de classificadores é um conjunto de classificadores que são utilizados para realizar uma predição conjunta (Imagem 1).



Fonte: Felipe Augusto (2020)

Alguns motivos para se usar comitês de classificadores:

- Diferentes classificadores com diferentes inputs: classificação por voz, características, imagens entre outros;
- Diferentes características, podem existir casos onde em um conjunto de dados falte algumas informações que em outros não e vice versa;
- A união faz a força, unir vários classificadores em uma mesma tarefa pode melhorar o desempenho da classificação.

Um comitê de classificadores tende a seguir uma configuração padrão

- Vários classificadores que avaliam o conjunto de dados separadamente, usando métricas, valores, inputs distintos ou iguais;
- Um combinador que irá agrupar esses valores, seguindo alguma técnica, para o exemplo foi se utilizado o somatório dos classificadores e o produto dos mesmos.

Implementação do algoritmo

Primeiramente foi necessário rodar o dataset Iris nos classificadores do weka, os classificadores escolhidos foram: J48 (Imagem 2), BayesNet (Imagem 3) e RandomForest (Imagem 4), com as configurações que o professor passou, porém foi percebido que no item 1.3.5 da especificação do trabalho dizia para habilitar o Output predictions, porém para essa versão do weka, essa opção não existia, ou esta com o nome trocado, utilizei a OutputDistribution (Imagem 5), creio que seja apenas um erro na descrição, ou então o fato de estar usando a versão de MacOs.

Imagem 2 – J48

The screenshot shows the Weka GUIChooser interface with the J48 classifier selected. The 'Test options' section is configured with 'Cross-validation' set to 'Folds 10'. The 'Result list' on the left shows several models, with '09.23.27 - treesJ48' selected. The 'Classifier output' pane displays the following data:

```
weka.classifiers.trees.J48 -C 0.25 -M 2
11,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.977,0.023
12,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.977,0.023
13,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.977,0.023
14,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.977,0.023
15,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.977,0.023
1,3:Iris-virginica,3:Iris-virginica,,0,0,0.024,0.976
2,3:Iris-virginica,3:Iris-virginica,,0,0,0.024,0.976
3,3:Iris-virginica,3:Iris-virginica,,0,0,0.024,0.976
4,3:Iris-virginica,3:Iris-virginica,,0,0,0.024,0.976
5,3:Iris-virginica,3:Iris-virginica,,0,0,0.024,0.976
6,1:Iris-setosa,1:Iris-setosa,,1,0,0
7,1:Iris-setosa,1:Iris-setosa,,1,0,0
8,1:Iris-setosa,1:Iris-setosa,,1,0,0
9,1:Iris-setosa,1:Iris-setosa,,1,0,0
10,1:Iris-setosa,1:Iris-setosa,,1,0,0
11,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.977,0.023
12,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.977,0.023
13,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.977,0.023
14,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.977,0.023
15,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.977,0.023

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      144      96 %
Incorrectly Classified Instances      5       4 %
Kappa statistic                    0.96
Mean absolute error                 0.025
Root mean squared error             0.1586
Relative absolute error              7.8785 %
Root relative squared error         31.6333 %
Total Number of Instances          158

=== Detailed Accuracy by Class ===
      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC   ROC Area  PRC Area  Class
-----
0.980   0.000   1.000   0.980   0.990   0.985   0.990   0.987   Iris-setosa
0.940   0.030   0.940   0.940   0.940   0.910   0.952   0.980   Iris-versicolor
0.960   0.030   0.941   0.960   0.950   0.925   0.961   0.985   Iris-virginica
Weighted Avg.   0.960   0.020   0.960   0.960   0.960   0.940   0.960   0.924

=== Confusion Matrix ===
  a b c  <-- classified as
40 1 0 | a = Iris-setosa
 0 47 3 | b = Iris-versicolor
 0 2 48 | c = Iris-virginica
```

Imagem 3 - BayesNet

The screenshot shows the Weka GUIChooser interface with the BayesNet classifier selected. The 'Test options' section is configured with 'Cross-validation' set to 'Folds 10'. The 'Result list' on the left shows several models, with '09.23.19 - bayesBayesNet' selected. The 'Classifier output' pane displays the following data:

```
weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5
11,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.999,0.001
12,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.999,0.001
13,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.999,0.001
14,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.999,0.001
15,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.999,0.001
1,3:Iris-virginica,3:Iris-virginica,,0,0,0.002,0.998
2,3:Iris-virginica,3:Iris-virginica,,0,0,0.002,0.998
3,3:Iris-virginica,3:Iris-virginica,,0,0,0.002,0.998
4,3:Iris-virginica,3:Iris-virginica,,0,0,0.002,0.998
5,3:Iris-virginica,3:Iris-virginica,,0,0,0.002,0.998
6,1:Iris-setosa,1:Iris-setosa,,1,0,0
7,1:Iris-setosa,1:Iris-setosa,,1,0,0
8,1:Iris-setosa,1:Iris-setosa,,1,0,0
9,1:Iris-setosa,1:Iris-setosa,,1,0,0
10,1:Iris-setosa,1:Iris-setosa,,1,0,0
11,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.999,0.001
12,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.999,0.001
13,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.999,0.001
14,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.999,0.001
15,2:Iris-versicolor,2:Iris-versicolor,,0,0,0.999,0.001

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      139      92.6667 %
Incorrectly Classified Instances      11      7.3333 %
Kappa statistic                    0.99
Mean absolute error                 0.004
Root mean squared error             0.1828
Relative absolute error             38.2111 %
Root relative squared error         38.7793 %
Total Number of Instances          158

=== Detailed Accuracy by Class ===
      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC   ROC Area  PRC Area  Class
-----
1.000   0.000   1.000   1.000   1.000   1.000   1.000   1.000   Iris-setosa
0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   Iris-versicolor
0.980   0.060   0.882   0.980   0.931   0.816   0.978   0.919   Iris-virginica
Weighted Avg.   0.927   0.037   0.927   0.927   0.927   0.890   0.908   0.942

=== Confusion Matrix ===
  a b c  <-- classified as
50 0 0 | a = Iris-setosa
 0 44 0 | b = Iris-versicolor
 0 5 45 | c = Iris-virginica
```

Imagem 4 - RandomForest

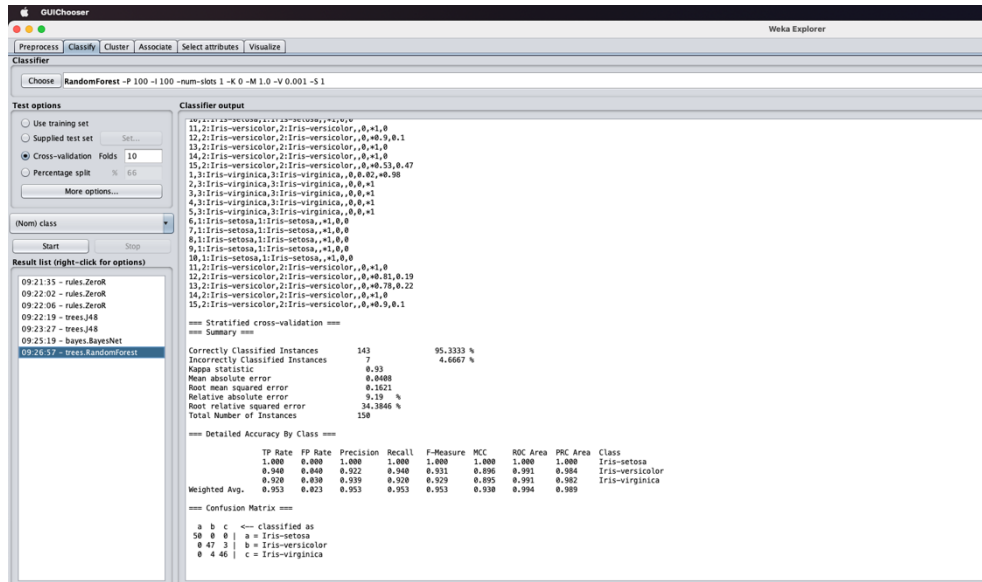
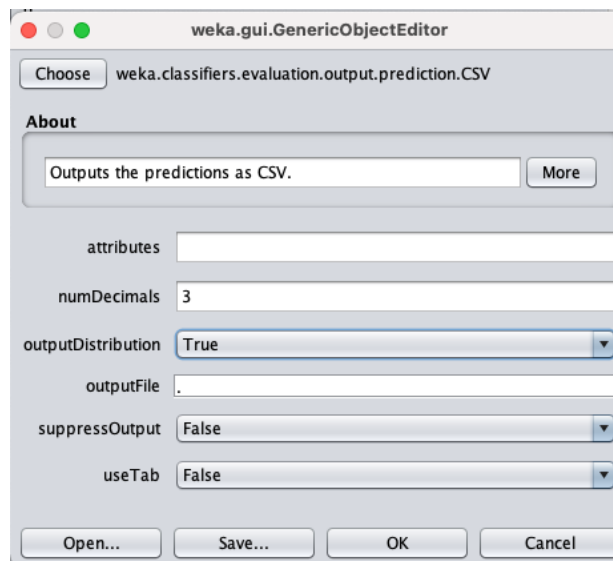


Imagem 5 - outputDistribution



Após rodados os classificadores e exportados os mesmos para CSV foi iniciado a implementação do código, inicialmente tentou-se se utilizar da biblioteca pandas para manipular os arquivos, porém por conta da falta de experiência com a mesma se optou em se utilizar apenas a csv, padrão do python.

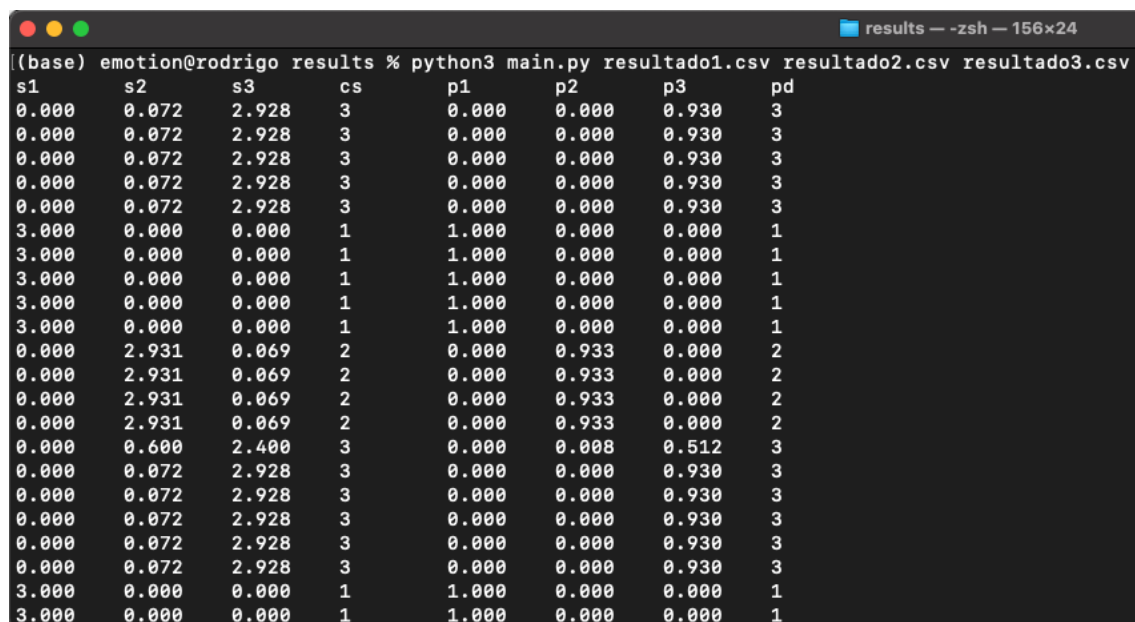
O código em si ficou bem simples, basicamente ele inicia carregando os arquivos, com base nos nomes passados via arguments: `python3 main.py resultado1.csv resultado2.csv resultado3.csv`

Com isso o algoritmo irá carregar os três arquivos, armazenar seus registros em listas, após isso o mesmo irá percorrer as listas realizando os cálculos de s1, s2, s3, cs, p1, p2 e pd. Como os datasets são iguais, o numero de registros em cada lista será o mesmo, logo foi tranquilo de implementar os cálculos. Ele realiza um somatório com os elementos das 3 listas para as variáveis correspondentes e exibe na tela.

Também foi criado um método chamado formatNumber, afim de formatar melhor as saídas.

A saída do algoritmo pode ser verificada na Imagem 6, e o código do mesmo se encontra no github¹.

Imagem 6 – Saída do algoritmo



| s1 | s2 | s3 | cs | p1 | p2 | p3 | pd |
|-------|-------|-------|----|-------|-------|-------|----|
| 0.000 | 0.072 | 2.928 | 3 | 0.000 | 0.000 | 0.930 | 3 |
| 0.000 | 0.072 | 2.928 | 3 | 0.000 | 0.000 | 0.930 | 3 |
| 0.000 | 0.072 | 2.928 | 3 | 0.000 | 0.000 | 0.930 | 3 |
| 0.000 | 0.072 | 2.928 | 3 | 0.000 | 0.000 | 0.930 | 3 |
| 0.000 | 0.072 | 2.928 | 3 | 0.000 | 0.000 | 0.930 | 3 |
| 3.000 | 0.000 | 0.000 | 1 | 1.000 | 0.000 | 0.000 | 1 |
| 3.000 | 0.000 | 0.000 | 1 | 1.000 | 0.000 | 0.000 | 1 |
| 3.000 | 0.000 | 0.000 | 1 | 1.000 | 0.000 | 0.000 | 1 |
| 3.000 | 0.000 | 0.000 | 1 | 1.000 | 0.000 | 0.000 | 1 |
| 3.000 | 0.000 | 0.000 | 1 | 1.000 | 0.000 | 0.000 | 1 |
| 0.000 | 2.931 | 0.069 | 2 | 0.000 | 0.933 | 0.000 | 2 |
| 0.000 | 2.931 | 0.069 | 2 | 0.000 | 0.933 | 0.000 | 2 |
| 0.000 | 2.931 | 0.069 | 2 | 0.000 | 0.933 | 0.000 | 2 |
| 0.000 | 2.931 | 0.069 | 2 | 0.000 | 0.933 | 0.000 | 2 |
| 0.000 | 0.600 | 2.400 | 3 | 0.000 | 0.008 | 0.512 | 3 |
| 0.000 | 0.072 | 2.928 | 3 | 0.000 | 0.000 | 0.930 | 3 |
| 0.000 | 0.072 | 2.928 | 3 | 0.000 | 0.000 | 0.930 | 3 |
| 0.000 | 0.072 | 2.928 | 3 | 0.000 | 0.000 | 0.930 | 3 |
| 0.000 | 0.072 | 2.928 | 3 | 0.000 | 0.000 | 0.930 | 3 |
| 0.000 | 0.072 | 2.928 | 3 | 0.000 | 0.000 | 0.930 | 3 |
| 3.000 | 0.000 | 0.000 | 1 | 1.000 | 0.000 | 0.000 | 1 |
| 3.000 | 0.000 | 0.000 | 1 | 1.000 | 0.000 | 0.000 | 1 |

¹ https://github.com/rodrigorafaeldamaceno/classifiers_committee

Referencias

Felipe Augusto, **Comitê de classificadores, Classificação multiclasse e multirrótulo. 2020**

Disponível em: <https://medium.com/@f2acode/comit%C3%AA-de-classificadores-classifica%C3%A7%C3%A3o-multiclasse-e-multirr%C3%B3tulo-7a01d87ee9b8>. Acesso em: 14/12/2021.