Data Mining - Ensemble Learning - Feature Selection Methods

Luisa Rosa - Spring 2024

Feature selection is used to remove irrelevant or correlated features in order to improve classification performance. In this project you can compare 2 different feature selection methods: the Filter Method which doesn't make use of cross-validation and the Wrapper Method which does.

Instructions:

- Download all files (4 Python programs and 1 .arff dataset)
- To see the Feature Selection Methods, run the respective python program
 - Run fm_from_scratch.py to see the Filter Method applied without any libraries
 - Run filter_method.py to understand the Filter Method
 - Run wrapper_method.py to understand the Wrapper Method
 - Run majority_vote.py to understand Ensemble Learning Majority Vote

_

Question 1: done.

TEAM 3:

Vanessa Arteaga, Luisa Rosa, Jonathan Ramos, Kayla Laufer

Working in Python.

Question 2: Filter Method

Make the class labels numeric (set "noncar"=0 and "car"=1) and calculate the Pearson Correlation Coefficient (PCC) of each feature with the numeric class label. The PCC value is commonly referred to as r.

(a) List the features from highest |r| (the absolute value of r) to lowest, along with their |r| values. Why would one be interested in the absolute value of r rather than the raw value?

(b) Select the features that have the highest m values of |r|, and run LOOCV on the dataset restricted to only those m features. Which value of m gives the highest LOOCV classification accuracy, and what is the value of this optimal accuracy?

Solution:

1. Step 1 - Convert labels and extract features and class labels into their own variables.

2. Step 2 - Calculate PCC for each feature.

3. Step 3 - Sort features in descending order based on their absolute correlation coefficients.

4. Step 4 - Print sorted features with their absolute correlation coefficients.

5. Step 5 - Fix k = 7 for all runs of Leave-One-Out Cross-Validation.

6. Step 6 - Perform Leave-One-Out-Cross-Validation for different values of m.
   - Select top m features based on their absolute correlation coefficients
   - Restrict the dataset to only those m features
   - Perform LOOCV
   - Update the best accuracy and corresponding m value

Answer:

2.a) Features from highest |r| to lowest and their |r| values:
Feature 4 |r| = 0.43692179751745097 r = 0.43692179751745097
Feature 13 |r| = 0.36826904080902556 r = 0.36826904080902556
Feature 14 |r| = 0.36822372149726035 r = -0.36822372149726035
Feature 16 |r| = 0.36602511423650724 r = 0.36602511423650724
Feature 7 |r| = 0.35214126136392476 r = 0.35214126136392476
Feature 22 |r| = 0.3513499255347562 r = 0.3513499255347562
Feature 26 |r| = 0.341042614938154 r = -0.341042614938154
Feature 1 |r| = 0.308810814577297 r = 0.308810814577297
Feature 20 |r| = 0.29904900743176754 r = 0.29904900743176754
Feature 31 |r| = 0.29078291134679674 r = 0.29078291134679674
Feature 34 |r| = 0.2660927897329379 r = 0.2660927897329379

```
Feature 2 |r| = 0.1957323905545247 r = 0.1957323905545247
Feature 28 |r| = 0.15690433267657294 r = 0.15690433267657294
Feature 25 |r| = 0.1530959899437491 r = 0.1530959899437491
Feature 19 |r| = 0.13763622058924627 r = 0.13763622058924627
Feature 17 |r| = 0.11394472976390747 r = 0.11394472976390747
Feature 32 |r| = 0.09317373256743067 r = 0.09317373256743067
Feature 8 |r| = 0.08777300962449965 r = -0.08777300962449965
Feature 0 |r| = 0.06979505192021568 r = -0.06979505192021568
Feature 10 |r| = 0.056876488922882024 r = 0.056876488922882024
Feature 21 |r| = 0.05660516824019749 r = -0.05660516824019749
Feature 11 |r| = 0.04211688398659001 r = -0.04211688398659001
Feature 33 |r| = 0.03880964782127297 r = 0.03880964782127297
Feature 6 |r| = 0.03529477533787 r = 0.03529477533787
Feature 15 |r| = 0.03147794480573731 r = -0.03147794480573731
Feature 35 |r| = 0.030855237164366982 r = 0.030855237164366982
Feature 29 |r| = 0.020829454125099515 r = 0.020829454125099515
Feature 18 |r| = 0.01793142539798069 r = -0.01793142539798069
Feature 27 |r| = 0.015606234757601551 r = -0.015606234757601551
Feature 9 |r| = 0.013005370998714192 r = -0.013005370998714192
Feature 3 |r| = 0.009213581453788188 r = 0.009213581453788188
Feature 30 |r| = 0.008955194801128104 r = 0.008955194801128104
Feature 24 |r| = 0.0077797430847418745 r = 0.0077797430847418745
Feature 23 |r| = 0.005507866072800373 r = 0.005507866072800373
Feature 12 |r| = 0.0021785840278091897 r = 0.0021785840278091897
Feature 5 |r| = 9.813778651783369e-05 r = 9.813778651783369e-05

By sorting based on |r| values, you can identify which features have the
strongest association with the class label, regardless of whether the
relationship is positive or negative. This helps in feature selection or
identifying important predictors in a dataset.

2.b)
Optimal value of m: 20
Highest LOOCV classification accuracy: 95.03546%

---
```

Question 3: Wrapper Method

Starting with the empty set of features, use a greedy approach to add the single feature that improves performance by the largest amount when added to the feature set. This is Sequential Forward Selection. Define performance as the LOOCV classification accuracy of the KNN classifier using only the features in the selection set (including the "candidate" feature). Stop adding features only when there is no candidate that when added to the selection set increases the LOOCV accuracy.

(a) Show the set of selected features at each step, as it grows from size zero to its final size (increasing in size by exactly one feature at each step)
(b) What is the LOOCV accuracy over the final set of selected features?

Solution:

1. Step 1 - Convert labels and extract features and class labels into their own variables.

2. Step 2 - Fix k = 7 for all runs of Leave-One-Out Cross-Validation.

3. Step 3 - Define variables and arrays to perform Sequential Forward Selection.

4. Step 4 - Loop through the feature list:
   - Combine selected features with the current feature
   - Select columns from the dataset corresponding to the selected features
   - Compute accuracy using LOOCV
   - Update selected feature subset if accuracy improves

5. Step 5 - Print the Selected Feature Subset in each iteration.
   - identifying the feature selected and the maximum accuracy achieved in that iteration

6. Step 6 - Determine the Final Features Selected and the final accuracy with the best feature set.

Answer:

3.a)

```
Step 0:
Selected feature subset is []
Step 1:
Maximum Accuracy achieved is 0.7423167848699763%, with feature f14
New Selected feature subset is ['f14']
Step 2:
Maximum Accuracy achieved is 0.8652482269503546%, with feature f10
New Selected feature subset is ['f14', 'f10']
Step 3:
Maximum Accuracy achieved is 0.9018912529550828%, with feature f19
New Selected feature subset is ['f14', 'f10', 'f19']
Step 4:
Maximum Accuracy achieved is 0.9361702127659575%, with feature f8
New Selected feature subset is ['f14', 'f10', 'f19', 'f8']
Step 5:
Maximum Accuracy achieved is 0.9562647754137116%, with feature f7
New Selected feature subset is ['f14', 'f10', 'f19', 'f8', 'f7']
Step 6:
Maximum Accuracy achieved is 0.958628841607565%, with feature f25
New Selected feature subset is ['f14', 'f10', 'f19', 'f8', 'f7', 'f25']
Step 7:
Maximum Accuracy achieved is 0.9621749408983451%, with feature f1
New Selected feature subset is ['f14', 'f10', 'f19', 'f8', 'f7', 'f25',
'f1']
Step 8:
Maximum Accuracy achieved is 0.9621749408983451%, with feature f9
New Selected feature subset is ['f14', 'f10', 'f19', 'f8', 'f7', 'f25',
'f1', 'f9']
Step 9:
Maximum Accuracy achieved is 0.9609929078014184%, with feature f13
Accuracy is not increased from the previous feature set. Breaking out of
loop.
```

3.b)
Final Selected Feature set is , ['f14', 'f10', 'f19', 'f8', 'f7', 'f25', 'f1', 'f9']
Final Accuracy with above feature set is 0.9621749408983451

Question 4: Ensemble Learning - Majority Vote
Suppose we need to build a predictive model for a binary classification task. We have 25 students in our class. Each of us worked independently and everyone is able to builda model with 60% accuracy.
a) If we take 3 models and build a majority vote classifier C3, what would be theaccuracy of our new classifier C3? Show your work.
b) If we take 5 models and build a majority vote classifier C5, what would be theaccuracy of our new classifier C5? Show your work.
c) If we take all 25 models and build a majority vote classifier C25, what would bethe accuracy of our new classifier C25? Show your work.
d) The performance you obtained for C25 is too good to be true. What's the as-sumption in your calculations that often does not hold in reality?
e) What would be the answer to (c) if everyone's model only has 45% accuracy? Show your work.

Answer:
4.a) Accuracy being 0.6 for 3 models: 0.648

4.b) Accuracy being 0.6 for 5 models: 0.6825600000000001

4.c) Accuracy being 0.6 for 25 models: 0.8462322310242371

4.d) In the majority vote ensemble method, it is assumed that each classifier in the ensemble makes predictions independently of the others. However, in practice, this assumption may not hold true. If the classifiers in the ensemble are trained on similar data or share common features, they may end up making correlated predictions. In such cases, the ensemble's performance may not be as good as expected based on the assumption of independence.

4.e) Accuracy being 0.45 for 25 models: 0.3063239659244824

3118/24

QUESTION 4 - Ensemble Learning - Luisa Rosa

4.a) Students: 25
Accuracy: 60%.

$\Rightarrow \sum_{k>\frac{n}{2}}^{n} \binom{n}{k} p^k (1-p)^{n-k}$

probability of correct prediction: 60%.
probability of wrong prediction: 40%.

→ ${}_3^2C + {}_3^3C$ ⇝ if classifier is correct

→ $\binom{3}{2}(0.6)^2(0.4)^1 = 3 \cdot 0.36 \cdot 0.4 = 0.432$
→ $\binom{3}{3}(0.6)^3(0.4)^0 = 1 \cdot 0.216 \cdot 1 = 0.261$

Correct classification rate: $0.432 + 0.261 = 0.648$
The accuracy of the majority vote classifier C3 is: 64.8%.

b) $C5 = {}_3^5C + {}_4^5C + {}_5^5C$
→ $\binom{5}{3}(0.6)^3(0.4)^2 = 10 \cdot 0.216 \cdot 0.16 = 0.3456$
→ $\binom{5}{4}(0.6)^4(0.4)^1 = 5 \cdot 0.1296 \cdot 0.4 = 0.2592$
→ $\binom{5}{5}(0.6)^5(0.4)^0 = 1 \cdot 0.07776 \cdot 1 = 0.07776$
Correct classification rate: $0.3456 + 0.2592 + 0.07776 = 0.68256$
the accuracy of C5 is: 68.256%.

c) 0.846232 → C25 accuracy ⇒ 84.6232%.
e) 0.30632 → when each accuracy is 45% ⇒ 30.632%.
d) In the majority vote ensemble method, it's assumed that each classifier in the ensemble makes predictions independently of the others. However, in practice, this assumption may not hold true. If the classifiers in the ensemble are trained on similar data or share common features, they may end up making correlated predictions. In such cases, the ensemble's performance may not be as good as expected based on the assumption of independence.